

Characterizing discourse group roles in inquiry-based university science labsTong Wan^{*}*Department of Physics, University of Central Florida,
4111 Libra Drive, Orlando, Florida 32816, USA*

Juliette Pimbert

*Department of Mechanical and Aerospace Engineering, University of Central Florida,
12760 Pegasus Drive, Orlando, Florida 32816, USA*

Reshawna L. Chapple

School of Social Work, University of Central Florida, 12805 Pegasus Drive, Orlando, Florida 32816, USA

Ying Cao

*School of Education and Child Development, Drury University,
900 North Benton Avenue, Springfield, Missouri 65802, USA*

Pierre-Philippe A. Ouimet

*Department of Physics, University of Regina,
3737 Wascana Parkway, Regina, Saskatchewan, S4S 0A2, Canada*

(Received 11 June 2024; accepted 18 June 2025; published 9 July 2025; corrected 4 March 2026)

Group work is commonly adopted in university science laboratories. However, student small-group discourse in university science labs is rarely investigated. We aim to bridge the gap in the literature by characterizing student discourse group roles in inquiry-based science labs. The instructional context for this study was a summer program hosted at a private research university in the eastern United States. The program was designed as a bridge program for matriculating students who were first generation and/or deaf or hard of hearing (DHH). Accommodations such as interpreters and technological tools were provided for DHH students. We analyzed 19 students' discourse moves in five lab activities from the video recordings, resulting in a total of 48 student-lab units. We developed codes to describe student discourse moves: *asking a question*, *proposing an idea*, *participating in discussion*, *chatting off-task*, and *talking with instructor*. Through a cluster analysis using the 48 student-lab units on quantified discourse moves, we identified four discourse styles, *high on-task high social*, *high on-task low social*, *low on-task high social*, and *low on-task low social*. The results show that individual students tended to demonstrate varying discourse styles in different lab activities; students' discourse styles within the same groups tended to be aligned with their group members. By examining group members' discourse styles in mixed-gender groups, we did not observe a difference in engagement levels between female and male students. DHH students in mixed hearing ability groups, however, were observed to have a lower level of engagement compared to their non-DHH group members. We discuss possible factors that may have contributed to the observations for genders and students with different hearing abilities. We also provide suggestions for promoting equitable small-group discourse in university science labs.

DOI: [10.1103/g4gf-w1yd](https://doi.org/10.1103/g4gf-w1yd)**I. INTRODUCTION****A. Group work in university science labs**

Group work is a typical format in university science laboratories, where students work in small groups to conduct experiments, collect data, and perform analysis. Students collaborate and build on each other's ideas to construct scientific knowledge, develop reasoning and experimentation skills, and practice communicating scientific ideas.

^{*}Contact author: tong.wan@ucf.edu

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

Because of its potential benefits, group work is commonly integrated with many student-centered instructional approaches that are implemented beyond the lab setting. These approaches include but are not limited to guided inquiry, peer instruction [1], and project-based learning [2]. Studies have shown that in a collaborative studio setting, close attention to group dynamics can help instructors provide intermittent support to get students unstuck and develop ideas with their peers [3–6]. Instructional strategies based on the instructor’s attention to students’ group interactions include making student thinking visible, noticing and re-voicing student ideas, supporting group interactions to include all students, and facilitating the sharing of resources [7–9].

Despite the overall improvement in student learning that group work may have contributed to, affective risks [10,11] and inequities [12–16] in group work have been documented in the literature. For example, Cooper *et al.* [11] found that students’ anxiety increases when they do not feel comfortable or they do not have a positive relationship with their work partners. Quinn *et al.* [13] reported inequitable gender division of labor in inquiry-based labs. Theobald *et al.* [14] found that students who perceived that they worked with a dominator had decreased performance. Dasgupta *et al.* showed that female students’ verbal participation increased as the proportion of women in small mixed-gender groups increased [16].

B. Student discourse in collaborative science learning

Although it is now a consensus that understanding group dynamics can increase learning and improve the learning experience, categorizing specific aspects of group dynamics is still developing. Prior research on students’ group roles using observation data in university science labs [13,15] tends to focus on student nonverbal behaviors (e.g., handling equipment, using computer, and writing on paper). Studies that explore student small-group discourse in university science lab settings are scarce.

Analyzing student discourse allows researchers to gain insights into how students interact with their group members and contribute to the group’s cognitive [9,17,18] and metacognitive development [19–21]. This is because cognition and metacognition processes are made visible through discourse. Cao and Korestky [9], for example, examined students’ discourse on shared resources in studios during interactive virtual labs on thermodynamics. They categorized conditions when sharing resources is likely to happen and found that interaction is necessary, and facilitating conditions include describing a discrepancy; requesting feedback; asking a general question; responding to a question by elaborating, rephrasing, or paraphrasing others’ ideas; making ideas public; and providing an alternative idea.

In addition to socially distributed cognitive discourse, researchers are also engaged in examining social

metacognitive discourse in small groups. As stated by Chiu and Kuo [20], social metacognition “distributes metacognitive demands among group members, it increases the visibility of one another’s metacognition, improves individual cognition, promotes reciprocal scaffolding, and enhances motivation.” Specifically in the field of physics education research, Sayre and Irving [19] identified a new element in the physicist speech genre, which they coined as brief, embedded, spontaneous metacognitive talk (BESM talk). In Sayre and Irving’s study, students use BESM talk to communicate their expectations in four ways: understanding, confusion, spotting inconsistencies, and generalized expectations.

We situate our study within the literature on students’ collaborative learning in small groups and contribute to this body of research by categorizing students’ group roles through the analysis of their discourse.

C. Current study

The overarching goal of this study is to characterize group roles students take on during discourse interaction in inquiry-based university science labs. The instructional context was a summer program for matriculated students, among whom a significant fraction of students were deaf or hard of hearing (DHH). Through video coding of coarse grain-sized student *discourse moves* (i.e., in-the-moment verbal actions) and clustering on those discourse moves, we aim to gain insights into how students engage in small-group discourse in labs. Specifically, we investigate the following research questions (RQs):

RQ1: How do the students in a summer program engage in small-group discourse in inquiry-based university science labs?

- a. What are the students’ small-group discourse styles in inquiry-based university science labs?
- b. Do the individual students tend to use the same or varied styles in different lab activities?
- c. Do the group members’ discourse styles tend to be aligned or diverse?

RQ2: Are students from historically underrepresented groups in a summer program engaged in equitable small-group discourse in inquiry-based university science labs?

- a. Are the female students engaged in equitable small-group discourse in inquiry-based university science labs?
- b. Are the DHH students engaged in equitable small-group discourse in inquiry-based university science labs?

II. METHODS

A. Instructional context

1. The IMPRESS program

The context for the current study was the Integrating Metacognitive Practices and Research to Ensure Student

Success (IMPRESS) [22] summer program hosted at a private research university in the eastern United States. It was a 10-day bridge program for matriculating students who were first generation and/or DHH. This program was offered yearly at the university from 2014 to 2017.

The main objectives of the IMPRESS program were to engage students in authentic science practice, to facilitate the development of a supportive community, and to help the students reflect on science and themselves in order to strengthen their learning habits and lead them to a stronger future in STEM fields. The program admitted up to 20 students each year. Students worked in small groups of three or four.

In the current study, we focus our analysis on laboratory experiments conducted in small groups during the time span of IMPRESS. Specifically, we analyzed five lab activities from the year 2015. We chose to focus on the year 2015 for three reasons. First, the video data from the years 2015 and 2016 were best documented and cataloged. Second, the ratio of DHH to non-DHH students in 2015 was almost one to one (9 to 11), which allowed us to make a comparison between the two groups of students. The DHH students in 2016, on the other hand, only made up a small fraction of the total participants (4 out of 19). Third, the communication modes used by DHH students in 2015 included both sign language and speech, but all the DHH students in 2016 used speech. The DHH students in 2015 were thus a better representation of the entire DHH student population.

2. Lab activities

What follows is a description of the lab activities in which the students participated as part of the program. We considered an activity a *lab* if it required experimental design or model construction, data collection, and data analysis. If an activity only required an observation of a phenomenon, it was not included in the analysis. Additionally, all the activities included occurred in an indoor lab setting; the outdoor activities were not included.

The overall goal of the IMPRESS lab activities is to help students gain a better understanding of the greenhouse effect—in other words, to see how increasing levels of CO₂

in the Earth's atmosphere “trap” heat. In lab 1, students were tasked with building a model of the Earth and its atmosphere. While building this initial model, students were given minimal instructions and allowed to use any of the materials provided. When the model had been constructed, the students then took it through “day” and “night” cycles using a heat lamp as a stand-in for the Sun. During this process, they took temperature measurements.

The students were also asked to describe what their expectations were for how the temperature of the model should behave given what they already knew about the greenhouse effect. The results of the temperature measurements were then compared to their expectations.

Following the initial model building, the students were taken through three tailored testing experiments in labs 2–4 that focused on different aspects of the ground-atmosphere system, namely thermal absorption, the albedo effect, and the greenhouse effect. Students tested the effects of different materials on the temperature of the model. When these activities had been completed, they were asked to build a new model of the ground-atmosphere system that incorporates what they had learned. In lab 5, students built an apparatus to collect and measure the amount of CO₂ released in combustion. A detailed description of the lab activities can be found in Appendix A.

The approximate duration, estimated time allocated for data collection, and the main objectives for each lab are shown in Table I. When students were not collecting data, they were either constructing their physical model or analyzing/interpreting the data collected. We note that students spent varying amounts of time collecting data (and also varying amounts of time building models and analyzing/interpreting data) in different labs, which could be a contributing factor for the types and/or amount of students' discourse moves.

B. Video data and participants

In the year 2015, 20 students participated in the program. Among all the participants, 10 were self-identified female and 9 were self-identified DHH, who were provided accommodations such as interpreters and computers. Students were divided into five groups,

TABLE I. Approximate duration, estimated time allotted for data collection, and main objectives of each lab.

	Approximate duration (min)	Estimated time allocated for data collection (min)	Main objectives
Lab 1	90	40	To build an initial model for Earth and Earth's atmosphere
Lab 2	120	120	To test thermal absorption
Lab 3	150	120	To test the Albedo effect
Lab 4	120	80	To test the greenhouse effect and build a new model for Earth and Earth's atmosphere
Lab 5	60	Not specified	To build an apparatus to collect and measure the amount of CO ₂ released in combustion

each with four members. The program activities were facilitated by a lead teacher and two teaching assistants (TAs), who were participants from the previous year. The teacher and TAs used questioning techniques to facilitate metacognitive discussion.

Most of the program activities, including the first four labs, took place in the IMPRESS classroom and the last lab was conducted in a chemistry laboratory. Students were self-grouped during the lab activities. They worked in the same groups in labs 1 and 2. At the beginning of lab 3, students were told to work with someone they had not worked with before. They were again self-grouped. The groups then remained the same for the rest of the labs.

A consent form was given to the students. All the students being recorded consented to participation in the research. All the recorded groups had all the members being part of the research.

For each lab, 3–4 groups were recorded by the cameras attached to the tables with a fixed angle, resulting in 17 recorded group-lab units. Five of the group-lab units were missing key information and therefore were not

coded. Two units were missing a significant fraction (about two-fifths) of the lab period; two were missing a group member in the recordings (a member was absent or not captured by the camera); in another unit where no interpreter was present, a DHH student had most of his body blocked in view by a computer, which made the signing not fully visible. We coded the groups only if all the members could be analyzed because we aim to characterize students' discourse group roles. If a member was missing from the analysis, the results would not be comparable to those of groups with all four members included.

The 12 units with complete information were coded. As a result, 19 of the 20 program participants were included in the analysis: 8 were DHH and 11 were non-DHH; 10 were female and 9 were male.

The group compositions with female and DHH students labeled are shown in Table II. Some of the DHH students used speech to communicate, and others used sign language with or without speaking simultaneously. For some groups, an interpreter was present at the table to help interpret when a DHH student signed. Two groups had

TABLE II. Group compositions with pseudonyms and the labs coded. Female (F) and DHH students with their modes of communication [23] are indicated in the parentheses.

Group composition		Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
Initial groups	Arya (F)	✓				
	Brittany (F)					
	Daniel					
	Pat (F, DHH, speaking)					
	Brock	✓	✓			
	Herb					
	Jakob					
	Justin					
	Adam (DHH, signing and simultaneous speaking)	✓				
	Ashley (F, DHH, signing)					
	Jessica (F, DHH, signing and speaking)					
	Taylor (F, DHH, signing)					
New groups	Brittany (F)			✓	✓	✓
	Jessica (F, DHH, signing and speaking)					
	Justin					
	Pat (F, DHH, speaking)					
	BJ (DHH, signing and simultaneous speaking)			✓	✓	
	Brett					
	Herb					
	Jakob					
	Ashley (F, DHH, signing)			✓	✓	
	Grace (F)					
	Jill (F, DHH, speaking)					
	Sara (F)					
	Arya (F)					✓
	Daniel					
	Jack (DHH, signing)					
	Tasha (F)					

times when no interpreter was present; author R. L. C., who is fluent in American Sign Language (ASL), transcribed students' sign language based on her interpretation. One of the coders (T. W.) and R. L. C. watched some of the videos together and discussed the transcript to make sure the coder interpreted the transcript accurately.

It is worth noting that the mixed-gender groups tended to have a majority of female members; 5 out of 6 mixed-gender units had 3 female members, and the sixth unit had an equal number of males and females. Of the six single-gendered units, four were all-male and two were all-female.

The DHH students in the mixed groups of hearing ability, on the other hand, tend to have slightly lower representations; four out of nine units had only one DHH student, and five units had an equal number of DHH and non-DHH students. Additionally, two units had no DHH students, and one unit had four DHH students.

C. Development of codes for discourse moves in labs

Author T. W. developed the initial definitions of the codes that describe students' discourse moves in science labs. Authors T. W. and Y. C. first independently coded 10-min worth of videos for all four students in a group, documented instances of discourse moves, and took detailed notes. They then compared their coding and resolved inconsistencies. Authors T. W., Y. C., and P. P. A. O. refined the code definitions and finalized the codes.

The finalized codes for discourse moves are shown in Table III. The codes for five different types of discourse moves include *asking a question*, *proposing an idea*, *participating in discussion*, *chatting off-task*, and *talking with the instructor*.

Inquiry-based labs often require students to make decisions throughout all phases. When we developed the codes, we intended to capture key discourse moves that help move the group forward toward shared group goals. In the following, we justify our development of codes based on the literature and our understanding of what is important in inquiry-based science labs. Moreover, the codes are grounded in our observations of the video data as they represent the themes that emerged from student discourse.

As pointed out by Goos *et al.* [21], "producing mutually acceptable solution methods and interpretations thus entails reciprocal interaction, which would require students to propose and defend their own ideas, and to ask their peers to clarify and justify any ideas they do not understand." Asking questions, proposing ideas, and participating in discussion are considered key discourse moves that contribute to the decision-making process. By asking a question, a student invites their group members to contribute, which helps make sure the group members come to a consensus. When a student proposes an experiment-related idea, the entire group is given an opportunity to critique and consider alternatives. When a question has been asked or an idea has been proposed, it is expected that the group members respond and engage in meaningful discussion. The code *participating in discussion* includes reciprocal interaction, such as answering questions, providing feedback on other members' ideas, as well as justifying one's own ideas, and making observations. We chose to focus on the coarse grain-sized discourse moves as an initial step toward characterizing group roles. We did not further differentiate how a student participates in the discussion because it is beyond the scope of the current study.

TABLE III. Code definitions and examples for student discourse moves in inquiry-based university labs.

Code	Definition	Example
Asking a question	Student asks group member a task-related, nonrhetorical question	"Now what are your ideas?" "How are we going to use this in our model?"
Proposing an idea	Student proposes an idea regarding experimental design/procedures, or data collection or analysis. It may or may not be in the form of a question. Note: when a question is posed, the coder first evaluates whether it concerns an idea about the experimental design or procedure. If it is, the coder codes "proposing an idea"; if not, the coder codes "asking a question" (when it is task related).	"Let's use this as a representation of land." "We can put holes in that and call it carbon dioxide." "Do we want to put the mirror inside the tank?"
Participating in discussion	Student participates in discussion, including answering questions, providing feedback to other's ideas, and rhetorical questions	"Because the mat is too big to fit in and there won't be room for ocean, so-" "So that doesn't retain heat but the black does."
Chatting off-task	Student talks with their group on topics not related to the task at hand, such as their personal life, jokes, and pronunciation. Can be a statement or a question.	"That takes some serious skill, drowning in the water." "That's why I'm not taking Bio."
Talking with instructor	Student talks with the lead teacher or a TA. It can be a statement or a question. It can be either on-task or off-task.	TA: "Where are you guys so far?" Student: "We have the sun, and we are going to put the saran wrap on top."

Chatting off-task can help group members build rapport with one another in a collaborative learning context. More importantly, prior research has shown several productive functions of off-task talks in relation to collaboration, such as “gaining access to collaboration for self,” “recruiting others into the collaboration,” and “resisting concentrated authority” [24–27].

Talking with the instructor may be an indicator of a student’s willingness to communicate with the instructor. Students could ask the instructor a question or discuss ideas with them. The interaction with an instructor could help the group move forward and also demonstrate a student’s confidence in their own knowledge and skills.

D. Video coding and interrater reliability

Typically, classroom observations of student and instructor behaviors (e.g., COPUS [28] and LOPUS [29]) were coded in 2-min intervals. We decided to reduce the interval duration to 1 min because, through an initial exploration of our video data, we noticed that a discourse move occurred more frequently than every 2 min. On the other hand, if the duration was too small, it would be very challenging for the coders to reliably capture all the discourse moves with the given quality of the video recordings. Therefore, we decided to code student discourse moves in 1-min interval. If a particular type of discourse moves for a student *ever* occurred within a 1-min interval, a result of 1 was documented; otherwise, a 0 was documented. We were able to achieve a reasonable level of interrater reliability (IRR) for coding every minute.

The videos were coded by authors T.W. and J.P. in Behavioral Observation Research Interactive Software (BORIS), an event logging software. To investigate IRR, authors T.W. and J.P. independently coded a total of 100-min worth of videos for all group members from two different groups for lab 1. We then compared the codes and obtained a Cohen’s kappa of 0.71, indicating a substantial agreement (i.e., 0.61–0.80). To explore code-level IRR, we used Gwet’s AC1 because Gwet’s AC1 is less sensitive to low trait prevalence compared to Cohen’s kappa [30]. In other words, Cohen’s Kappa does not accurately reflect the level of agreement when a code has low prevalence, which may be the case for certain codes (e.g., *proposing an idea*). The Gwet’s AC1 we obtained for *asking a question*, *proposing an idea*, *participating in discussion*, *chatting off-task*, and *talking with the instructor* are 0.89, 0.87, 0.61, 0.72, and 0.87, respectively. The agreement on each code falls in the range of substantial or almost perfect agreement (i.e., 0.81–1.00).

E. Cluster analysis

To characterize student discourse styles, we conducted a cluster analysis using the five discourse move variables in R. Through clustering, objects that are similar to one another are grouped in the same cluster. The object of our

analysis is a student’s profile, defined by the prevalence of the five discourse variables in a lab activity. Since we coded 12 group-lab units and each group had 4 students, the number of student profiles was 48.

To prepare data for a cluster analysis, we first determined the frequency (i.e., number of 1-min intervals) that each variable was coded. To account for different durations of the lab activities, we divided the frequency by the total number of minutes that the activity lasted for.

We used Ward’s method, a hierarchical agglomerative clustering technique. This type of clustering technique works in a bottom-up manner. Each student profile was initially a cluster. The profiles that were most similar (i.e., closest in distance in a five-dimensional space determined by the five variables) were combined into a bigger cluster. This was repeated until all the profiles were combined in a single cluster. To determine the optimal number of clusters, we used the “elbow method” [31]. We found the “elbow” at cluster number $k = 5$ (see Appendix B). The follow-up statistical analysis showed that two of the five clusters were not disguisable from each other. We then combined the two clusters, resulting in a total of four clusters. The four clusters were significantly different from one another in terms of all five variables with large effects (see Sec. III).

F. Statistical analysis

To examine the ways in which the clusters are different, we first used Kruskal-Wallis rank sum test [32]. The effect size was calculated using eta-squared [33] with $0.01 \leq \eta^2 < 0.06$ being small effect, $0.06 \leq \eta^2 < 0.14$ being moderate effect, and $\eta^2 \geq 0.14$ being large effect [34]. We then conducted a *post hoc* analysis using Dunn’s multiple comparison test [35] with Holm-Bonferroni corrections. All the statistical analyses were conducted in R.

Due to a small sample size, we did not conduct statistical tests to examine the correlation among gender, hearing ability, and discourse moves. Instead, we used a measure, which we named the “comparative engagement level” index to compare discourse engagement between gender groups, and between DHH and non-DHH students, respectively, while controlling other variables (see Secs. III D and III E for detail).

III. RESULTS

The distribution of fractions of 1-min intervals for all student profiles for each discourse code is shown in Fig. 1. The three most prevalent but also widespread codes are *participating in discussion* ($M = 0.35$, $SD = 0.21$), *chatting off-task* ($M = 0.25$, $SD = 0.19$), and *talking with instructor* ($M = 0.20$, $SD = 0.12$). The other two types of discourse moves *asking a question* ($M = 0.10$, $SD = 0.07$) and *proposing an idea* ($M = 0.08$, $SD = 0.09$) occurred a lot less frequently.

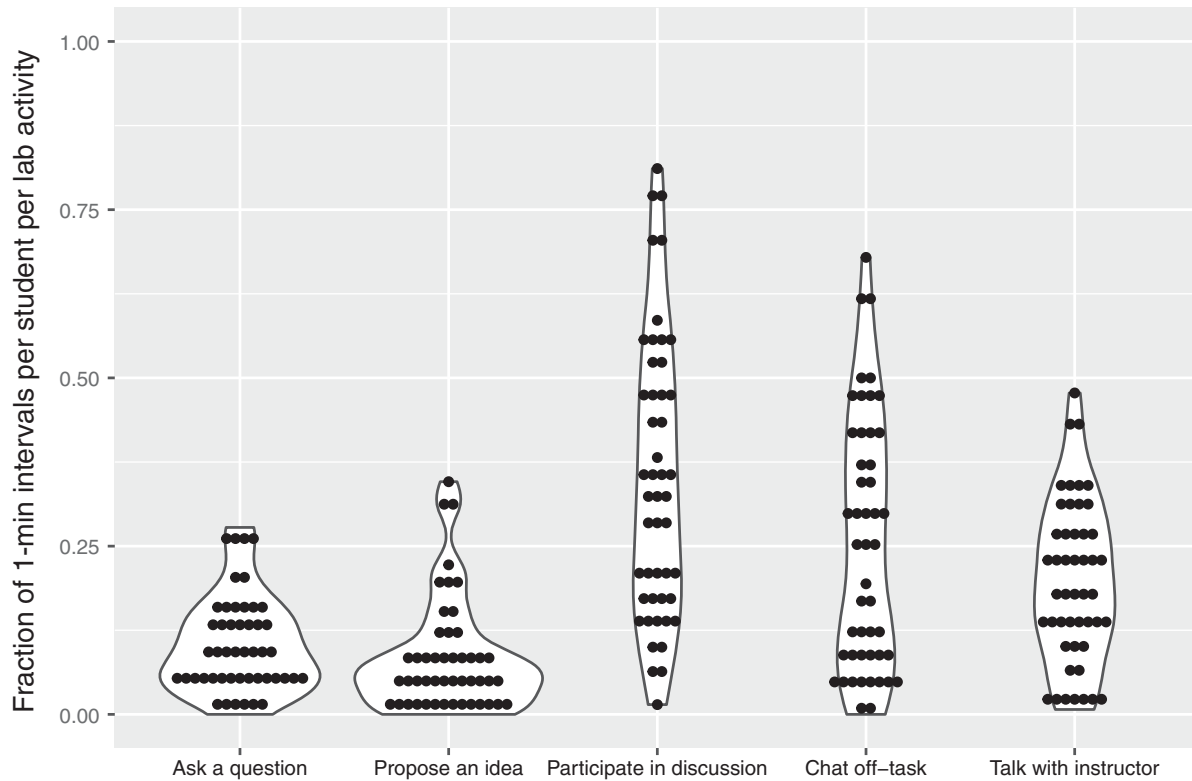


FIG. 1. Distribution of fractions of 1-min intervals for all student profiles for each discourse code. Data points represent the student profiles ($N = 48$) generated from the video coding.

A. RQ1a: Characteristics of student discourse styles

The hierarchical structure and cluster memberships of the student discourse profiles are shown in Fig. 2. The numbers of profiles in clusters A, B, C, and D are 14, 11, 9, and 14, respectively.

The distribution of fractions of 1-min intervals for each discourse code in each cluster is shown in Fig. 3. The Kruskal-Wallis tests showed that at least one of the four clusters is distinct from the others in terms of all five discourse codes ($p = 0.002$ for *asking a question*, and $p < 0.001$ for the others) with large effects. Therefore, we use all five discourse codes to describe the characteristics of the four clusters of students' discourse styles. From here on, we refer to the clusters as student discourse styles.

Table IV shows a summary of the characteristics of student discourse styles. The ranking of the clusters for each code was determined by Dunn's test. For the code *asking a question*, for example, the ranking of the clusters is $A = C > B = D$. Thus, clusters A and C are considered to have high fractions of asking a question, and clusters B and D are considered to have low fractions (see Appendix C, Table X for complete results).

Interestingly, the ways that the four styles are different from one another are identical for the three codes that describe on-task collaboration, *asking a question*, *proposing an idea*, and *participating in discussion*. That is,

styles A and C had higher fractions for these three codes compared to styles B and D. Therefore, we grouped the first three codes into a “on-task” category. Moreover, the code *chatting off-task* can be used to describe the extent to which students are socializing with their group members. Recall that prior research has suggested that “social” (or off-task talk) has productive functions for collaborative learning.

Since four out of the five codes describe the interaction among group members, the four codes (or the two categories, on-task and social) should characterize the main features of the clusters. Therefore, we describe the main characteristics of the four discourse styles in terms of the extent to which students are engaging in on-task collaboration and socializing with their group members. Additionally, we discuss the frequency of talking with the instructor for each discourse style.

Style A: High on-task low social. Students who used the high on-task low social style asked more questions, proposed more ideas, and participated more often in discussions. They chatted off-task less frequently and had a medium level of interaction with the instructor or TA.

Style B: Low on-task low social. Students who used the low on-task low social style not only engaged in on-task collaboration less often, but they also chatted with their group and talked with instructors less often.

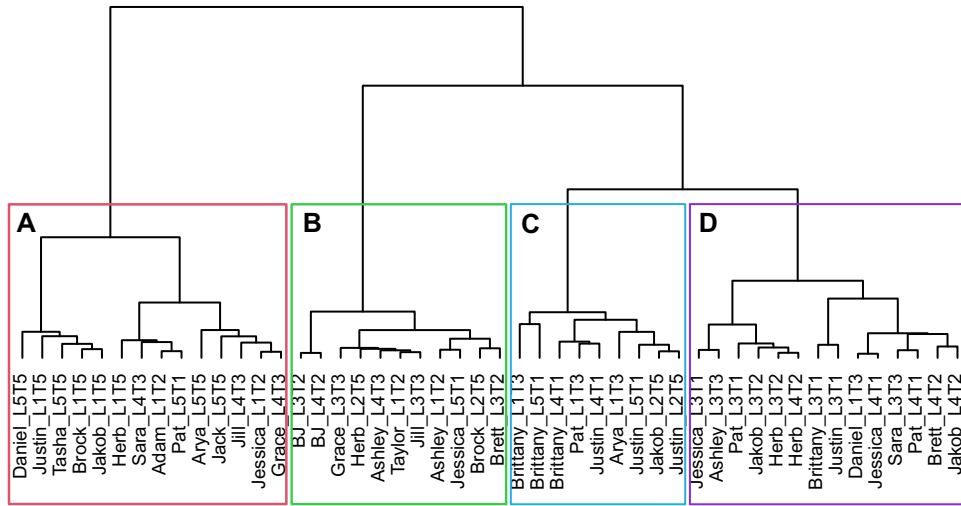


FIG. 2. Dendrogram showing the hierarchical structure of the student discourse profiles. Each box represents a cluster of student discourse profiles, each labeled by the student's pseudonym, the lab number, and the table (also the group) number.

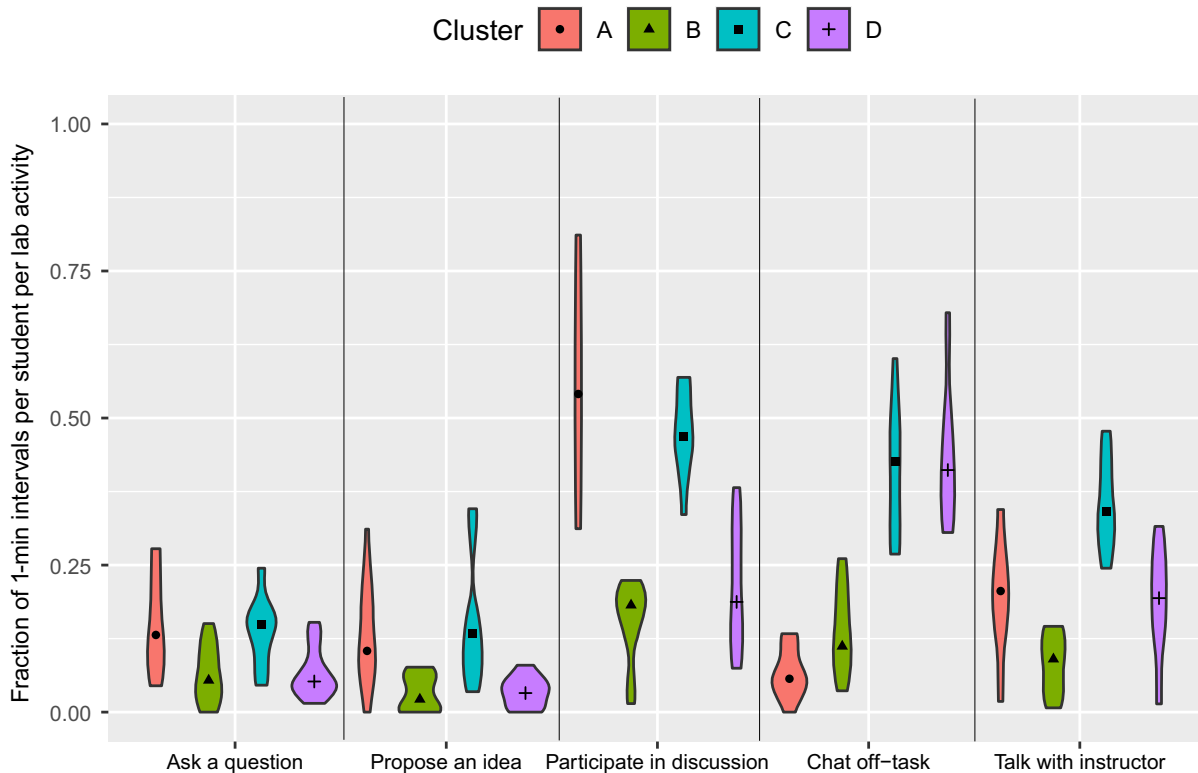


FIG. 3. Distribution of fractions of 1-min intervals for each discourse code in each cluster. Points represent median fractions.

Style C: High on-task high social. Students who used the high on-task high social style asked more questions, proposed more ideas, and participated more often in discussions. They also chatted off-task more often and talked with instructors most frequently.

Style D: Low on-task high social. Students who used the low on-task high social style engaged in on-task

collaboration less often, but they chatted more often with their group and had a medium level of interaction with the instructor.

It is worth noting that style C (high on-task high social) is the most desirable among all four styles because it has a high engagement level for all five discourse codes. The second most desirable style is A (high on-task low social),

TABLE IV. Characteristics of student discourse styles in each cluster. Median fractions of 1-min intervals are shown in parentheses.

Cluster label	Discourse style	Asking a question ($A = C > B = D$, Dunn's)	Proposing an idea ($A = C > B = D$, Dunn's)	Participating in discussion ($A = C > B = D$, Dunn's)	Chatting off-task ($A = B < C = D$, Dunn's)	Talking with instructor ($B < A = D < C$, Dunn's)
A	High on-task low social	High (0.13)	High (0.10)	High (0.54)	Low (0.06)	Medium (0.21)
B	Low on-task low social	Low (0.05)	Low (0.02)	Low (0.18)	Low (0.11)	Low (0.09)
C	High on-task high social	High (0.15)	High (0.13)	High (0.47)	High (0.43)	High (0.34)
D	Low on-task high social	Low (0.05)	Low (0.03)	Low (0.19)	High (0.41)	Medium (0.19)

followed by style D (low on-task high social). The least desirable style is B (low on-task low social), which has a low engagement level for all five codes.

B. RQ1b: Individual students' discourse styles

We found that individual students tended to use a variety of discourse styles in different labs. As shown in Table V, the number of labs in which an individual student was coded ranged from 1 to 5. For students who were coded in more than one lab, 14 out of 15 students used at least two discourse styles. An individual student could use up to three discourse styles.

TABLE V. Distribution of numbers of discourse styles used by individual students.

Number of labs coded for an individual student	Number of styles observed	Number of students
1	1	4
2	1	1
	2	7
3	2	1
	2	1
4	3	4
5	3	1

Students' use of a variety of discourse styles appears to be associated with the lab activities they worked on (see Table VI). Two or three groups were coded in all the labs except for lab 2, where only one group (four male non-DHH students) was coded. Here we discuss the distributions of the discourse styles for labs 1, 3–5. Students' on-task discourse moves started with high prevalences from lab 1 (50% for style A: high on-task low social and 25% for style C: high on-task high social) and then waned down in lab 3 (no student used style A or C with high on-task), and finally increased in lab 5 (63% for style A: high on-task low social and 25% for style C: high on-task high social).

Recall that students constructed their initial model of the Earth and Earth's atmosphere in lab 1 and then they tested the effects of new materials in labs 2–4. In lab 4, students also constructed a new model based on the results from all previous testing experiments. During the testing experiments, students repeatedly recorded the temperature for long periods of time (see Table I). This may be why they used the low on-task styles more prevalently in lab 3. Additionally, the objective of lab 5 was to build an apparatus to collect and measure the amount of CO₂ from burning material. In this lab, students probably spent more time on designing the experiment than collecting data, which may have contributed to a greater prevalence of the high on-task styles.

TABLE VI. Number of each discourse style observed in labs 1 and 3–5.

	Number of students coded	Number of each style observed			
		A: High on-task Low social	B: Low on-task Low social	C: High on-task High social	D: Low on-task High social
Lab 1	12	6 (50%)	2 (17%)	3 (25%)	1 (8%)
Lab 3	12	0	4 (33%)	0	8 (67%)
Lab 4	12	3 (25%)	2 (17%)	2 (17%)	5 (41%)
Lab 5	8	5 (63%)	1 (12%)	2 (25%)	0

TABLE VII. Frequency of number of styles observed within a group.

Number of styles within a group	Frequency	Observed styles in a group
1	3	AAAA, AAAA, DDDD
2	8	AAAB, AABB, BBCC, BBDD, BBDD, BDDD, CCCD, CCDD
3	1	ABCC

Moreover, students regrouped in lab 3 and then they stayed in the same groups for the rest of the program. Students were asked to work with someone whom they had never worked with before. This may be a reason why style D (low on-task high social) was much more prevalent than style B (low on-task low social) in lab 3. Students may have chatted off-task more frequently to socialize with the new groups, which could help them gain access to collaboration.

C. RQ1c: Group members' discourse styles

A student in a group often shared the same discourse style with at least another member. Out of all 12 group-lab units coded, 11 units had only one or two discourse styles observed (see Table VII). There were three instances where all members used the same styles.

Moreover, all four students' discourse styles within a group often shared at least one feature (i.e., high on-task shared by styles A and C, low on-task shared by styles B and D, high social shared by styles C and D, low social shared by styles A and B). For example, the group with styles AAAB shared low social across all the members, but the members with style A (high on-task low social) had a higher engagement in on-task discourse. There were two exceptions observed: the group with styles ABCC and the group with styles BBCC have no feature shared across all four members. The group with style ABCC, however, had a feature shared with two or three of the members in this group (A and B share low social, and A and C share high on-task).

D. RQ2a: Gender vs discourse

The distribution of fractions of 1-min intervals for each code for different genders for all 48 student profiles is shown in Fig. 4. The distributions appear very similar for female and male students.

To explore the correlation between gender and students' use of discourse styles, we examine individual mixed-gender lab-group units and compare the discourse styles of male and female students with the same hearing ability. By examining individual mixed-gender lab-group units, we can control the variables of lab activity and group. Since students worked in groups on the same task, it is more appropriate to compare students' discourse styles within

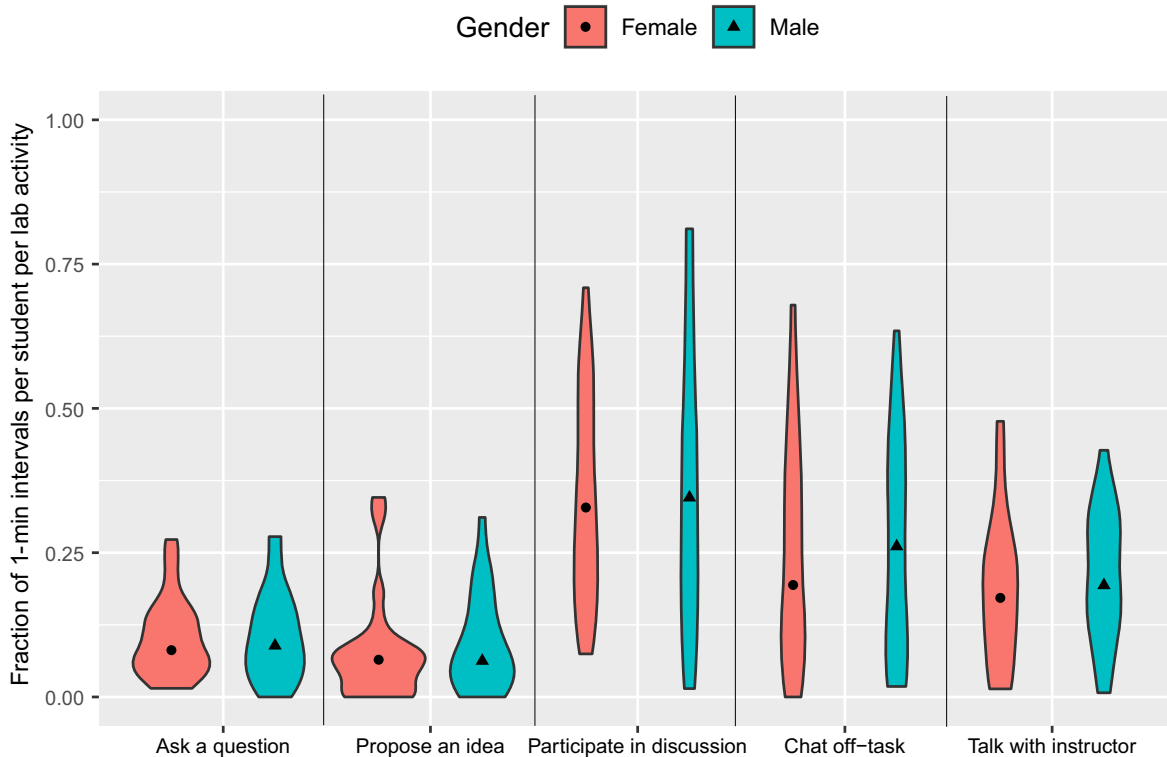


FIG. 4. Distribution of fractions of 1-min intervals for each code for different genders. Points represent median fractions.

groups. Additionally, we control the variable of hearing ability by comparing male and female students with the same hearing ability.

To quantify the comparison in engagement level between genders, we developed a metric, which we name the comparative engagement level (c.e.l.) index. This metric quantifies the normalized frequency that group members in one gender have a higher (or lower) engagement level than group members in another gender have. The index ranges from -1 to 1 , with a positive value meaning a higher engagement level and a negative value meaning a lower engagement level. After calculating the c.e.l. index for each lab-group unit, an average was calculated over all the mixed-gender lab-group units, n (see equation below).

$$\text{c.e.l.}_{\text{avg}} = \frac{1}{n} \sum_i^n \frac{\text{raw c.e.l. score}_i}{\text{max raw c.e.l. score}_i}$$

The maximum value of 1 means that a student in one gender, on average, has a higher engagement level than all other group members with a different gender; a value of 0.5 , for example, means that a student in one gender,

on average, has a higher engagement level than half of the other group members with a different gender.

To illustrate how the index was calculated for male students in mixed-gender units, we show the raw and maximum c.e.l. scores as well as the c.e.l. index for individual units in Table VIII. Recall that style C (high-on task high social) has the highest engagement level, followed by style A (high on-task low social), and then style D (low on-task high social), with style B (low on-task low social) having the lowest engagement level (see Sec. III A for detail). For simplicity, we use $C > A > D > B$ to represent the ranking from highest to lowest level of engagement. In unit L1T2, for example, all four members are DHH, and we compare the male student's (Adam's) discourse style with the other three female students' styles. Adam's style (style A: high on-task low social) has a higher engagement level than the two female (Ashley and Taylor) students' styles (both are style B: low on-task low social) and the same as one female student (Jessica). Adam gets a raw c.e.l. score of $+2$ because his style has a higher engagement level than that of two students of a different gender. The maximum raw c.e.l. score Adam could get is 3

TABLE VIII. Raw and maximum raw c.e.l. scores as well the c.e.l. index for male students in individual mixed-gender lab-group units. Each male student in a group was compared with every other female student with the same hearing ability. The ranking of the discourse styles from the highest to lowest engagement level is $C > A > D > B$. *Students not included in the comparisons due to different hearing abilities.

Unit label	Pseudonym	Gender	Hearing ability	Discourse style	Raw c.e.l. score for male	Max raw c.e.l. score for male	c.e.l. index for male in each unit
L1T2	Ashley	Female	DHH	B			
	Taylor	Female	DHH	B			
	Jessica	Female	DHH	A			
	Adam	Male	DHH	A	+2	3	+2/3
L1T3	Arya	Female	Non-DHH	C			
	Brittany	Female	Non-DHH	C			
	Daniel	Male	Non-DHH	D	−2	2	−1
	Pat*	Female	DHH	C			
L3T1	Brittany	Female	Non-DHH	D			
	Justin	Male	Non-DHH	D	0	1	0
	Jessica*	Female	DHH	D			
	Pat*	Female	DHH	D			
L4T1	Brittany	Female	Non-DHH	C			
	Justin	Male	Non-DHH	C	0	1	0
	Jessica*	Female	DHH	D			
	Pat*	Female	DHH	D			
L5T1	Brittany	Female	Non-DHH	C			
	Justin	Male	Non-DHH	C	0	1	0
	Jessica*	Female	DHH	B			
	Pat*	Female	DHH	A			
L5T5	Arya	Female	Non-DHH	A			
	Tasha	Female	Non-DHH	A			
	Daniel	Male	Non-DHH	A	0	2	0
	Jack*	Male	DHH	A			
Average c.e.l. index for male students per mixed-gender unit:							−0.06

because there are three students of a different gender with the same hearing ability. In some other units, there are only one or two female students with the same hearing ability, then the maximum is 1 or 2, respectively.

We also point out that one could calculate the c.e.l. index for the female students instead. The total raw c.e.l. for all three female students would have the same absolute value as the male student in the group but with a different sign. For example, in unit L1T2, the total raw c.e.l. scores for all three females is -2 (Ashely gets -1 , Taylor gets -1 , and Jessica gets 0). The total maximum raw c.e.l. for female students is the same as that for male students (with a maximum of 1 per student, the total maximum for three female students is 3). Because the numbers of male and female students with the same hearing ability in a group can be different, an average over all the mixed-gender units rather than over all the female (or male) students should be calculated to ensure consistent results for female and male students.

The average c.e.l. index for male students per mixed-gender unit is -0.06 , which is very close to zero. To further help interpret this value, we compare -0.06 to the average for the situation where a male student had a lower engagement level than *one* female student in the same group. We use $-1/3$ as a reference value for the following reasons. If a male student, on average, had a lower engagement level than one female student in the same group, the average

c.e.l. index for male students must have an absolute value that is greater than $1/3$. This is because, in one of the mixed-gender units, there are one male student and three female students with the same hearing ability (the raw c.e.l. score would be -1 , the maximum would be 3, and index $= -1/3$), and in the other mixed-gender units, there are one male and one or two female students (the raw c.e.l. score would be -1 , the maximum would be 1 or 2, and index $= -1$ or $-1/2$). Because -0.06 has an absolute value that is significantly smaller than the absolute value of $-1/3$, we conclude that no difference in engagement level was observed between male and female students in mixed-gender groups. Moreover, in four of the six mixed-gender units, the male student had a zero raw c.e.l. score. That is, we observed no difference between genders at both individual group level and overall.

E. RQ2b: Hearing ability vs discourse

The distribution of fractions of 1-min intervals for each code for students with different hearing abilities for all 48 student profiles is shown in Fig. 5. Overall, DHH students tend to have lower fractions across all codes compared to non-DHH students.

To control lab activity, group, and gender, we examine individual mixed-hearing-ability lab-group units and compare DHH and non-DHH students with the same gender. Similarly, we calculate the c.e.l. index for DHH students.

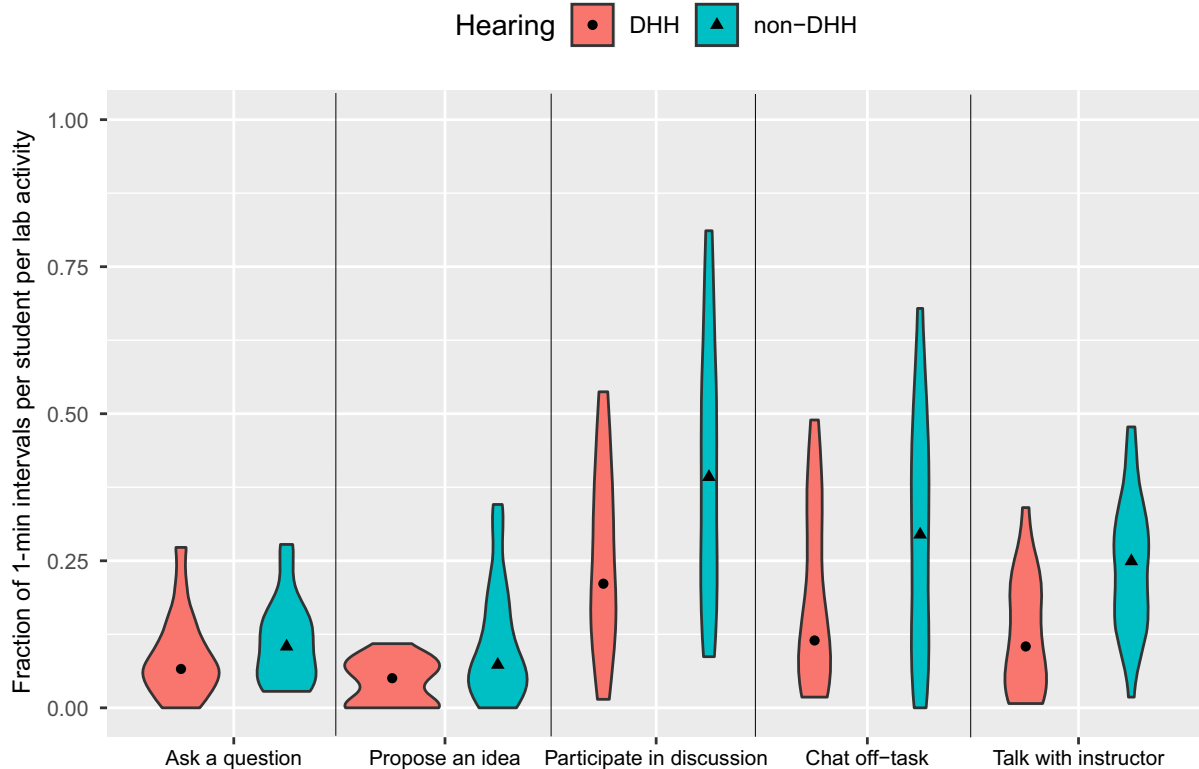


FIG. 5. Distribution of fractions of 1-min intervals for each code for students with different hearing abilities. Points represent median fractions.

TABLE IX. Raw and maximum raw c.e.l. scores as well the c.e.l. index for DHH students in individual mixed-hearing-ability lab-group units. DHH student in a group was compared with every other non-DHH student with the same gender. The ranking of the discourse styles from the highest to lowest engagement level is $C > A > D > B$. *Students not included in the comparisons due to different genders.

Unit label	Pseudonym	Gender	Hearing ability	Discourse style	Raw c.e.l. score for each DHH	Max raw c.e.l. score for each DHH student	c.e.l. index for DHH students in each unit
L1T3	Arya	Female	Non-DHH	C			
	Brittany	Female	Non-DHH	C			
	Pat	Female	DHH	C	0	2	0
	Daniel*	Male	Non-DHH	D			
L3T2	Brett	Male	Non-DHH	B			
	Herb	Male	Non-DHH	D			
	Jakob	Male	Non-DHH	D			
	BJ	Male	DHH	B	-2	3	-2/3
L4T2	Brett	Male	Non-DHH	D			
	Herb	Male	Non-DHH	D			
	Jakob	Male	Non-DHH	D			
	BJ	Male	DHH	B	-3	3	-1
L3T3	Grace	Female	Non-DHH	B			
	Sara	Female	Non-DHH	D			
	Ashley	Female	DHH	D	1	2	0
	Jill	Female	DHH	B	-1	2	
L4T3	Grace	Female	Non-DHH	A			
	Sara	Female	Non-DHH	A			
	Ashley	Female	DHH	B	-2	2	-1/2
	Jill	Female	DHH	A	0	2	
L3T1	Brittany	Female	Non-DHH	D			
	Jessica	Female	DHH	D	0	1	0
	Pat	Female	DHH	D	0	1	
	Justin*	Male	Non-DHH	D			
L4T1	Brittany	Female	Non-DHH	C			
	Jessica	Female	DHH	D	-1	1	-1
	Pat	Female	DHH	D	-1	1	
	Justin*	Male	Non-DHH	C			
L5T1	Brittany	Female	Non-DHH	C			
	Jessica	Female	DHH	B	-1	1	-1
	Pat	Female	DHH	A	-1	1	
	Justin*	Male	Non-DHH	C			
L5T5	Daniel	Male	Non-DHH	A			
	Jack	Male	DHH	A	0	1	0
	Arya*	Female	Non-DHH	A			
	Tasha*	Female	Non-DHH	A			
Average c.e.l. index for DHH students per mixed-hearing-ability unit:							-0.46

As shown in Table IX, in five out of nine mixed-hearing-ability units, there are two DHH students within the group. For those five units, a total raw c.e.l. score for DHH students in each unit and a total maximum c.e.l. score in each unit was used to calculate the c.e.l. index.

The average c.e.l. index for DHH students per mixed-hearing-ability unit is -0.46 . The average is very close to -0.5 , which means a DHH student on average has a lower level of engagement than about half of the non-DHH students of the same gender in the same group. Moreover,

there was only one instance of a DHH student using style C (high-on task high social), the most desirable style, and that there were a disproportionate number of DHH students who used style B (low on-task low social), the least desirable style. The results show that the DHH students had a lower level of engagement compared to their non-DHH group members of the same gender.

In five (L3T2, L4T2, L4T3, L4T1, and L5T1) out of nine mixed units, DHH students in the groups had lower engagement levels compared to their non-DHH group

members. In the other four units, DHH and non-DHH students had an equal engagement level. By examining the five units where a difference between DHH and non-DHH students was observed, we found that DHH students appeared to have a lower level of engagement for on-task collaborations or social interactions compared to non-DHH students within the groups. For example, BJ in unit L4T2 used the low on-task low social (style B) while the other members used the low on-task high social (style D). Additionally, DHH students also tended to talk with the instructor or TA less often given that the high on-task high social (style C) had the most interactions with the instructor or TA and the low on-task low social (style B) had the least interactions.

IV. DISCUSSION

A. Interpretation of findings and implications

The students in the IMPRESS summer program were observed to demonstrate different discourse styles in different lab activities. When the lab provided ample opportunities for decision making (e.g., a significant portion of time on experimental design was required), more students demonstrated the two high on-task styles (with high prevalences of asking questions, proposing ideas, and participating in discussion). In contrast, when the lab did not require much decision-making (e.g., a significant portion of time on repeated measurements was required), more students demonstrated the two low on-task styles. This was consistent with findings by Quinn *et al.* [13] who found that students in inquiry-based labs engaged in a variety of activities (e.g., worked on desktop or laptop, handled equipment, and wrote on paper), while students in traditional labs mainly worked on the worksheets. Our findings confirm that the design of lab tasks plays an important role in prompting student engagement.

Moreover, students' use of discourse styles was also influenced by their group members. The results showed that some members in a group, if not all, usually shared the same discourse styles. Additionally, the discourse styles students demonstrated often had common features with those of their group members. Since students' discourse moves tended to be aligned with their group members, we suggest that interventions for promoting productive group work should focus on the group level rather than the individual level. Students should be given opportunities to reflect and evaluate collectively as a group on how well the group has been functioning. This is in line with the intervention used by Dew *et al.* [15] who tasked students, as a group, to complete partner agreement and reflection forms in addition to individual reflections.

We did not observe a difference in engagement level in discourse between female and male students in mixed-gender groups. We hypothesize that two potential factors may have contributed to this observation. First, most (five

out of six) of the mixed-gender groups had female students in the majority and one group had equal male and female students. Prior research has shown that women's verbal participation in small groups increased as the proportion of women increased [16]. Moreover, for first year students, women in female-majority or female-parity groups were less anxious than those in female-minority groups [16]. Another potential factor is that the female and male ratio in the 2015 IMPRESS program was one to one, which was drastically different from many of the university science classrooms (e.g., physics and computer science) where male students usually make up the majority. Studies have shown that the female-to-male ratio in a classroom is positively correlated with female in-class participation [36,37]. Findings from the studies mentioned above and the current study together can be interpreted using the "stereotype inoculation model" proposed by Dasgupta [38]. Female peers in class, and in groups particularly, can act like "social vaccines" to "inoculate" against stereotype threat and increase female students' sense of belonging in a group (and class). We suggest that female-majority groups should be prioritized when instructors assign student groups in science classrooms where female students are typically underrepresented.

By examining individual mixed-hearing-ability groups and controlling lab activity and gender, we found that DHH students had a lower level of engagement compared to their non-DHH group members. DHH students tended to participate less in on-task conversations and social interactions; they also tended to interact with the instructors less often. The results were consistent with a prior study in high school classrooms that DHH had lower participation than their non-DHH peers despite the availability of technological tools and sign language interpretation [39].

For the DHH students who had a lower level of engagement compared to their non-DHH peers, two students used speech to communicate, one used sign language, and the other two used both sign language and speech. There was no clear evidence that the mode of communication played a role, but it could be due to the small sample size.

One challenge DHH students appeared to face in the lab was that the room could get really loud with all the groups talking simultaneously. We would like to remind the reader that DHH students include those who are hard of hearing, which is why noise levels could affect them more than hearing students. Moreover, some group members were observed to have side conversations, which put an additional barrier for DHH students even when an interpreter was present because the simultaneous conversations could make interpretation very challenging. A prior study in the context of undergraduate student research also reported that a DHH was distracted during group meetings when too many conversations were happening simultaneously [40]. One way that could help mitigate this barrier is to raise

instructors' and non-DHH students' awareness of DHH students' discourse needs in a lab setting.

Another challenge faced by DHH students may have been confusion masked by the appearance of low motivation, disinterest in the activity, or boredom. In our data, there was one group where all four members (Adam, Ashley, Jessica, and Taylor) were DHH, and they used sign language to communicate. Ashley and Taylor (both had style B: low on-task low social) participated a lot less than Adam and Jessica (both had style A: high on-task low social) did. The low participation of Ashley and Taylor may have resulted from low motivation. During the lab, Ashley expressed that "I am so done with this" and "this is boring." Similarly, Taylor wrote in her reflective journal, "I had fun to play with foil, but this experiment is so boring, I didn't like it or have fun while do it ...". Moreover, Ashley appeared lost in the lab as she wrote in her journal, "There were really nothing for me to do since I had no clue what the heck we were doing by creating a model of the atmosphere. I was like what were we supposed to do with this?" Only about half of the journals scanned were identifiable, and thus we were not able to analyze students' journals to see how prevalent low motivation was among students with different hearing abilities (or with different genders). Nonetheless, we suggest instructors should increase efforts in promoting students' motivation, such as showing support for students and expressing interest in students' ideas and experiences. This would help increase students' sense of belonging, which could in turn help increase student participation.

B. Limitations and future work

This study has a limitation in its sample size. Of the 17 group-lab units, 5 were excluded from the analysis because they were missing key information (either a significant fraction of the lab period or a group member) in the recordings. Additionally, due to the limitations in the recordings, not every student was coded in an equal number of lab activities. Each lab was not coded an equal number of times either. However, our sample had a similar female-to-male ratio (10 to 9) to the entire program (10 to 10). The DHH to non-DHH ratio (8 to 11) was also similar to that in the entire program (9 to 11).

Moreover, this study used coarse grain-sized discourse variables to characterize students' discourse styles in university science labs. This allowed us to code the videos relatively efficiently while focusing on the most prominent features in students' discourse styles. However, when coding students' discourse moves, we observed nuanced differences in some of the codes. Analysis addressing those nuances is beyond the scope of this study but can be conducted in future work. For example, the code *participating in discussion* can be broken down into *answering a question*, *making an observation*, *providing feedback to others*, and *elaborating on one's idea*. Additionally, a time-sequence log of codes for each group member can

be used to examine group dynamics in greater detail. Furthermore, students' discourse moves can be analyzed and characterized at varying levels (type of student interaction, primary intent, and nature of utterance) using the analytical framework proposed by Nennig *et al.* [41].

The current study investigated students' small-group discourse styles in a science summer program and no gender gap was observed. Future research should investigate discourse styles in different genders in typical university science labs and test the effect of gender ratio in small groups on students' discourse.

Finally, another future research avenue, in the long term, is to evaluate the effectiveness of group-level interventions in promoting equitable discourse for DHH students. An example of such interventions can include the same elements that were reported in Dew *et al.* [15], an individual reflection, a partner agreement form, and a partner reflection form. However, these forms need to be modified to reflect discourse moves since the original version primarily focuses on what students are handling rather than discourse. Moreover, the challenges faced by DHH students in lab settings should be taken into consideration when developing these interventions.

V. CONCLUSION

In this study, we identified four discourse styles in students' group work in university science labs. Students' use of discourse styles was influenced by both the lab activities and their group members. We did not observe gender inequity. However, DHH students were found to have a lower level of engagement in task-related and social discourse, and they talked with the instructor or TA less often compared to their non-DHH peers. Suggestions have been made to promote equitable small-group discourse in university science labs.

ACKNOWLEDGMENTS

We thank the Professional Development for Emerging Education Researchers (PEER) Institute for making the video recordings available for analysis. We also thank Scott Franklin, Eleanor Sayre, and Mary Bridget Kustusich for their support and useful discussions. Last but not least, we thank the two anonymous referees for their insightful feedback.

APPENDIX A: DETAILED DESCRIPTION OF LAB ACTIVITIES

1. Lab 1: Initial model construction

On the morning of day 2, students constructed their initial model of the Earth and Earth's atmosphere using the equipment provided. They measured the temperature of their model during the heating and cooling processes. The students were given minimal guidance and were allowed to

proceed as they desired. Following the data collection, students compared the results with their expectations.

2. Lab 2: Thermal absorption

On the morning of day 3, students measured the temperature in a model (of the Earth and Earth’s atmosphere) with and without a black mat present. The goal was to explore how the heat absorption of a black mat affects the microclimate of the model.

3. Lab 3: Albedo effect

On the morning of day 4, students measured the temperature in a model (of the Earth and Earth’s atmosphere) with and without a white foam mat present. The goal was to investigate how the reflection of solar radiation influences the microclimate of the model.

4. Lab 4: Greenhouse effect—thermal absorption of CO₂

On the morning of day 5, students measured the temperatures of liquids in two flasks, one is pure water and the other contains dissolving tablets that would produce CO₂. The goal was to look at how the heating and cooling of the atmosphere differs with and without CO₂ in the air. Students were also tasked to revise their model of Earth and Earth’s atmosphere incorporating what they had learned from the first four labs.

5. Lab 5: Burning model

On the morning of day 7, students carried out a chemistry experiment. The purpose was to build an apparatus that would collect and measure the amount of CO₂ released from burning material.

APPENDIX B: ELBOW PLOT

The elbow plot shows the elbow at cluster number $k = 5$ (see Fig. 6).

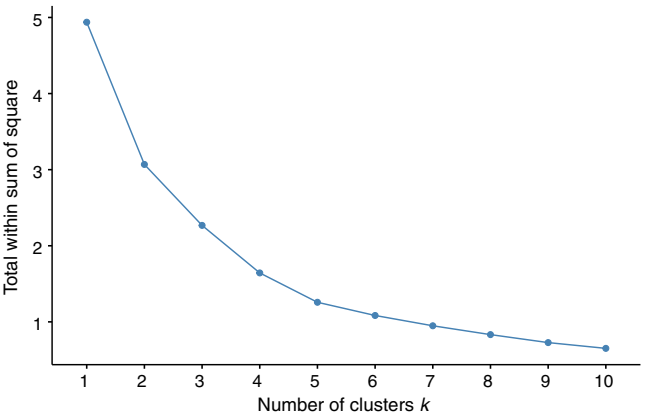


FIG. 6. Elbow plot used to determine the optimal number of clusters. The total within the sum of square is plotted as a function of the number of clusters.

APPENDIX C: DETAILED RESULTS OF KRUSKAL-WALLIS AND DUNN’S TESTS

TABLE X. Kruskal-Wallis and Dunn’s test results for each discourse code. Results of eta-squared for effect size are also shown. ** $p < 0.01$. *** $p < 0.001$. †Large effect.

Code	$\chi^2(3)$	p_{k-w}	η^2	Ranking by Dunn’s test
Asking a question	15.306	0.002**	0.280†	A = C > B = D
Proposing an idea	23.45	<0.001***	0.465†	A = C > B = D
Participating in discussion	33.262	<0.001***	0.688†	A = C > B = D
Chatting off-task	36.982	<0.001***	0.772†	A = B < C = D
Talking with instructor	25.274	<0.001***	0.506†	C > A = D > B

[1] C. H. Crouch and E. Mazur, Peer instruction: Ten years of experience and results, *Am. J. Phys.* **69**, 970 (2001).

[2] D. Kokotsaki, V. Menzies, and A. Wiggins, Project-based learning: A review of the literature, *Improv. Sch.* **19**, 267 (2016).

[3] M. Koretsky, J. Keeler, J. Ivanovitch, and Y. Cao, The role of pedagogical tools in active learning: A case for sense-making, *Int. J. STEM. Educ.* **5**, 18 (2018).

[4] M. D. Koretsky, Cognitive and social aspects of engagement in active learning, *Chem. Eng. Educ.* **51**, 198 (2017).

[5] M. D. Koretsky, D. Montfort, S. B. Nolen, M. Bothwell, S. C. Davis, and J. D. Sweeney, Towards a stronger covalent bond: Pedagogical change for inclusivity and equity, *Chem. Eng. Educ.* **52**, 117 (2018).

[6] M. D. Koretsky, Program level curriculum reform at scale: Using studios to flip the classroom, *Chem. Eng. Educ.* **49**, 47 (2015).

[7] I. S. Horn, *Strength in Numbers: Collaborative Learning in Secondary Mathematics* (National Council of Teachers of Mathematics, Reston, VA, 2012).

[8] M. Windschitl and A. Calabrese Barton, Rigor and equity by design: Locating a set of core teaching practices for the

- science education community, in *Handbook of Research on Teaching*, edited by D. H. Gitomer, and C. A. Bell, 5th ed. (AERA Press, Washington, DC, 2016), pp. 1099–1158.
- [9] Y. Cao and M. D. Koretsky, Shared resources: Engineering students' emerging group understanding of thermodynamic work, *J. Eng. Educ.* **107**, 656 (2018).
 - [10] L. D. Conlin and R. E. Scherr, Making space to sensemake: Epistemic distancing in small group physics discussions, *Cognit. Instr.* **36**, 396 (2018).
 - [11] K. M. Cooper, V. R. Downing, and S. E. Brownell, The influence of active learning practices on student anxiety in large-enrollment college science classrooms, *Int. J. STEM Educ.* **5**, 23 (2018).
 - [12] D. Doucette and C. Singh, Making lab group work equitable and inclusive, *J. Coll. Sci. Teach.* **52**, 31 (2023).
 - [13] K. N. Quinn, M. M. Kelley, K. L. McGill, E. M. Smith, Z. Whipps, and N. G. Holmes, Group roles in unstructured labs show inequitable gender divide, *Phys. Rev. Phys. Educ. Res.* **16**, 010129 (2020).
 - [14] E. J. Theobald, S. L. Eddy, D. Z. Grunspan, B. L. Wiggins, and A. J. Crowe, Student perception of group dynamics predicts individual performance: Comfort and equity matter, *PLoS One* **12**, e0181336 (2017).
 - [15] M. Dew, E. Hunt, V. Perera, J. Perry, G. Ponti, and A. Loveridge, Group dynamics in inquiry-based labs: Gender inequities and the efficacy of partner agreements, *Phys. Rev. Phys. Educ. Res.* **20**, 010121 (2024).
 - [16] N. Dasgupta, M. M. M. Scircle, and M. Hunsinger, Female peers in small work groups enhance women's motivation, verbal participation, and career aspirations in engineering, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 4988 (2015).
 - [17] J. Osborne, Arguing to learn in science: The role of collaborative, critical discourse, *Science* **328**, 463 (2010).
 - [18] S. Kaartinen and K. Kumpulainen, Collaborative inquiry and the construction of explanations in the learning of science, *Learn. Instr.* **12**, 189 (2002).
 - [19] E. C. Sayre and P. W. Irving, Brief, embedded, spontaneous metacognitive talk indicates thinking like a physicist, *Phys. Rev. ST Phys. Educ. Res.* **11**, 020121 (2015).
 - [20] M. M. Chiu and S. W. Kuo, Social metacognition in groups: Benefits, difficulties, learning, and teaching, *J. Educ. Res.* **3**, 1 (2009).
 - [21] M. Goos, P. Galbraith, and P. Renshaw, Socially mediated metacognition: Creating collaborative zones of proximal development in small group problem solving, *Educ. Stud. Math.* **49**, 193 (2002).
 - [22] S. V. Franklin, E. Hane, M. B. Kustus, C. Ptak, and E. C. Sayre, Improving retention through metacognition: A program for deaf/hard-of-hearing and first-generation STEM college students, *J. Coll. Sci. Teach.* **48**, 21 (2018).
 - [23] The DHH students in the study communicated using various communication modalities: speech, American Sign Language (ASL), Signing Exact English (SEE), the combination of spoken-English and sign language (Simultaneous communication or Sim-Com) or spoken English. DHH student's language use is directly tied to their Deaf cultural identity. For example, students who consider themselves culturally deaf through family or attendance at deafness-orientated schools will be more drawn to using ASL without speech. Students with non-DHH family members and friends may be drawn to using SEE signs, Sim-Com, or spoken-English. We do not include the specific sign languages students used in the table because we did not code students' signing directly; instead, we coded the interpretations by the interpreters.
 - [24] J. M. Langer-Osuna, E. C. Gargroetzi, R. Chavez, and J. Munson, Rethinking loafers: Understanding the productive functions of off-task talk during collaborative mathematics problem-solving, in *Proceedings of the 13th International Conference of the Learning Sciences (ICLS)* (2018), pp. 745–751.
 - [25] C. A. Hass, F. Genz, M. B. Kustus, P.-P. A. Ouimet, K. Pomian, E. C. Sayre, and J. P. Zwolak, Studying community development: A network analytical approach, in *Presented at PER Conference 2018, Washington, DC*, [10.1119/perc.2018.pr.Hass](https://doi.org/10.1119/perc.2018.pr.Hass).
 - [26] P. Hutchison, Equity and off-task discussion in a collaborative small group, in *Presented at PER Conf. 2022, Grand Rapids, MI*, [10.1119/perc.2022.pr.Hutchison](https://doi.org/10.1119/perc.2022.pr.Hutchison).
 - [27] J. M. Langer-Osuna, Productive disruptions: Rethinking the role of off-task interactions in collaborative mathematics learning, *Educ. Sci.* **8**, 87 (2018).
 - [28] 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).
 - [29] J. B. Velasco, A. Knedeisen, D. Xue, T. L. Vickrey, M. Abebe, and M. Stains, Characterizing instructional practices in the laboratory: The laboratory observation protocol for undergraduate STEM, *J. Chem. Educ.* **93**, 1191 (2016).
 - [30] N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet, A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples, *BMC Med. Res. Methodol.* **13**, 61 (2013).
 - [31] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, Integration K-means clustering method and elbow method for identification of the best customer profile cluster, *IOP Conf. Ser.* **336**, 012017 (2018).
 - [32] W. H. Kruskal and W. A. Wallis, Use of ranks in one-criterion variance analysis, *J. Am. Stat. Assoc.* **47**, 583 (1952).
 - [33] M. Tomczak and E. Tomczak, The need to report effect size estimates revisited. An overview of some recommended measures of effect size, *Trends Sport Sci.* **1**, 19 (2014).
 - [34] J. M. Maher, J. C. Markey, and D. Ebert-May, The other half of the story: Effect size analysis in quantitative research, *CBE Life Sci. Educ.* **12**, 345 (2013).
 - [35] O. J. Dunn, Multiple comparisons using rank sums, *Technometrics* **6**, 241 (1964).
 - [36] K. Canada and R. Pringle, The role of gender in college classroom interactions: A social context approach, *Sociol. Educ.* **68**, 161 (1995).
 - [37] 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**, 1 (2020).

- [38] N. Dasgupta, Ingroup experts and peers as social vaccines who inoculate the self-concept: The stereotype inoculation model, *Psychol. Inquiry* **22**, 231 (2011).
- [39] M. Salehomoum, Inclusion of signing deaf or hard-of-hearing students: Factors that facilitate versus challenge access and participation, *Perspect. ASHA Spec. Interest Groups* **5**, 971 (2020).
- [40] A. U. Gehret, J. W. Trussell, and L. V Michel, Approaching undergraduate research with students who are deaf and hard-of-hearing, *J. Sci. Educ.* **20**, 20 (2017).
- [41] H. T. Nennig, N. E. States, M. T. Montgomery, S. G. Spurgeon, and R. S. Cole, Student interaction discourse moves: Characterizing and visualizing student discourse patterns, *Discip. Interdiscip. Sci. Educ. Res.* **5**, 2 (2023).

Correction: A minor error in an inline equation following the third paragraph of Sec. III D and a missing value in the preceding sentence has been fixed.