CBE—Life Sciences Education
Vol. 14, 1–14, Winter 2015

Article

Cues Matter: Learning Assistants Influence Introductory Biology Student Interactions during Clicker-Question Discussions

Jennifer K. Knight,* Sarah B. Wise,† Jeremy Rentsch,*‡ and Erin M. Furtak§

*Department of Molecular, Cellular, and Developmental Biology, †Department of Ecology and Evolutionary Biology, and §School of Education, University of Colorado Boulder, Boulder, CO 80309

Submitted April 13, 2015; Revised August 15, 2015; Accepted August 17, 2015
Monitoring Editor: Jennifer Momsen

The cues undergraduate biology instructors provide to students before discussions of clicker questions have previously been shown to influence student discussion. We further explored how student discussions were influenced by interactions with learning assistants (LAs, or peer coaches). We recorded and transcribed 140 clicker-question discussions in an introductory molecular biology course and coded them for features such as the use of reasoning and types of questions asked. Students who did not interact with LAs had discussions that were similar in most ways to students who did interact with LAs. When students interacted with LAs, the only significant changes in their discussions were the use of more questioning and more time spent in discussion. However, when individual LA–student interactions were examined within discussions, different LA prompts were found to generate specific student responses: question prompts promoted student use of reasoning, while students usually stopped their discussions when LAs explained reasons for answers. These results demonstrate that LA prompts directly influence student interactions during in-class discussions. Because clicker discussions can encourage student articulation of reasoning, instructors and LAs should focus on how to effectively implement questioning techniques rather than providing explanations.

INTRODUCTION

One of the scientific practices most often emphasized in science education reform is argumentation, or the discussion and defense of competing ideas. Prior research has established that engaging students in argumentation can build students' abilities to understand, practice, and participate in science (Osborne *et al.*, 2004). In addition, argumentation has been shown to encourage scientific thinking, since this process involves students confronting different ideas about content as they describe their reasoning to one another (e.g., Kuhn, 1993; Koslowski, 1996; Zohar and Nehmet, 2002; Asterhan and Schwarz, 2009). Ultimately, students who are taught the principles of argumentation as part of their science courses have the potential to perform better on assessments that require reasoning (Bao *et al.*, 2009; Osborne, 2010). In fact, many recent publications about science education have emphasized the importance of engaging students in such practices to support their learning of science content (e.g., National Research Council [NRC], 2007; American Association for the Advancement of Science [AAAS], 2011). This focus is an extension of decades of efforts at the K–12 level that have sought to bring authentic science practices, critical thinking, and problem solving into classrooms and to engage students in scientific practices intertwined with their learning of scientific ideas (Duschl, 2008; NRC, 2012).

Unfortunately, despite national efforts, incorporating these practices has proven challenging for K–12 teachers

CBE Life Sci Educ December 1, 2015 14:ar41
DOI:10.1187/cbe.15-04-0093
‡Present address: Biology Department, Francis Marion University, Florence, SC 29502.
Address correspondence to: Jennifer K. Knight (jennifer.knight@colorado.edu).

(Windschitl, 2004; Bricker and Bell, 2008; Furtak and Alonzo, 2010) as well as for undergraduate instructors (Handelsman *et al.*, 2004; AAAS, 2011; President's Council of Advisors on Science and Technology, 2012). A number of studies have shown that wide-scale change is slow, even when instructors have been exposed to best practices (Henderson and Dancy, 2008; Dancy and Henderson, 2010). In traditional undergraduate science courses that can enroll hundreds of students, instructors often struggle to effectively engage students in argumentation practices as part of their course work, citing lack of time and resources to develop and implement changes.

One active-learning approach that is relatively easy to implement and often has dramatic effects on student engagement is the use of clicker questions with peer discussion (Mazur, 1997). Well-constructed clicker questions can provide students with challenging scenarios, thus encouraging students to argue about the possible reasons for their answers. Furthermore, engaging in clicker-question discussion positively impacts student performance in class, demonstrating that students are learning from their discussions (Smith *et al.*, 2009, 2011). Perhaps even more importantly, the process of interacting with peers who may have different ideas or ways of explaining content can stimulate students to engage in the process of constructing or reconstructing knowledge, which sociocognitive conflict theory suggests will improve learning (Asterhan, 2013).

One critical component to consider in student discussions is the effect that instructors may have on students through the behaviors they use to engender a certain class culture. Instructors vary in the ways they frame the goals of clicker use to students, the degree of interaction they have with students during the clicker discussion period, the time they allow for discussion, the extent to which they allow students to contribute ideas to whole-class discussion, whether both correct and incorrect answers are discussed during this summary, and the way that clicker participation is graded (Turpen and Finkelstein, 2009). Such instructional practices generate classroom norms that vary along a continuum, from placing emphasis on quickly arriving at the correct answer ("authoritative" or "answer-making") to placing emphasis on students articulating the reasons for both right and wrong answers ("dialogic" or "sense-making"; Turpen and Finkelstein, 2010). This variation impacts the way students perceive the goals of the instructor and the classroom norms, which in turn impacts the way students interact during clicker discussions (Turpen and Finkelstein, 2010). Indeed, the quality of student argumentation has been shown by others to depend on both the instructional approach and the scaffolding provided by the teacher (Songer and Gotwals, 2012). One study of fourth- and fifth-grade students working together to answer questions on photosynthesis demonstrated that written prompts that stimulated discussion (such as "Explain why your answer is correct or wrong" or "Can you compare how you used to think about this with how you think about it now?") elicited more structured discussion and further exploration of ideas than scenarios in which students were unprompted. The prompting also improved their performance on posttest questions above the improvement seen with unprompted students (Coleman, 1998). In addition, in a previous study on how specific types of cues affect student discussion (Knight *et al.*, 2013), we found that advanced students engaging in group discussions of clicker questions were sensitive to the language used by the instructor to cue the beginning of their discussion. When students were prompted to use reasoning and be prepared to discuss their reasoning with the rest of the class, they were significantly more likely to engage in the exchange of complete reasoning statements than when they were prompted to simply discuss the right answer. This finding strongly suggests that the students were paying attention to the cue given to them and were changing their interactions with one another dependent on this cue.

A second approach that may ameliorate challenges to encouraging argumentation in science undergraduate courses is the use of peer coaches to facilitate student in-class discussion. Peer coaches can encourage the use of reasoning and argumentation during class and can also guide students in problem solving and other hands-on activities. At the University of Colorado, peer coaches are called learning assistants (LAs), and the learning assistant program has become an international model for training students to serve in this role, primarily in large-enrollment introductory science courses (Otero, 2006). LAs are trained through a pedagogy course in the School of Education that emphasizes techniques to encourage students to construct their own ideas. In physics courses, LAs have been shown to improve student learning gains on concept assessments, and to improve student attitudes toward science (Otero *et al.*, 2010). Students in similar programs, such as peer-led team learning and peer-led guided inquiry, also benefit from interactions with their peer leaders, making significantly higher gains in science critical-thinking skills than their peers (Lewis and Lewis, 2008; Quitadamo *et al.*, 2009). Thus, the peer-coach model may be a way to encourage, model, and support argumentation in class, helping students to develop these skills while they are learning content.

In this paper, we kept constant how instructors cued the students and instead investigated how the presence of LAs might additionally impact student interactions during clicker discussions. We hypothesized that introductory students who were being cued by their instructors to use reasoning in their discussions would regularly use both questioning and reasoning but that the quality and quantity of such interactions would increase in the presence of a peer coach. We first characterize the general features of introductory students' clicker discussions, only some of which were conducted in the presence of a peer coach. We then characterize the discussions of groups who both did and did not interact with peer coaches and, finally, look more carefully at individual student responses to different kinds of cues from the peer coaches.

METHODS

Course Characteristics

We performed this study in a freshman-level introductory molecular and cell biology course that is required for students planning to major in this discipline and is also taken by students intending other majors (e.g., integrative physiology and neuroscience). Two experienced instructors (not the authors) cotaught two sections of this course, with a total enrollment of ~450 students; because of room and time

constraints, one section was large, and one was much smaller (94 students). This study was carried out in the smaller section of 94 students. The instructors spent class time (50 min, three times per week) engaged in lecture interspersed with three to five clicker questions per class period, using the iClicker response system, with time given for discussion and feedback on each question. The following clicker-question cycle with peer discussion (Mazur, 1997) was used most frequently: the instructor displayed the clicker question and asked students to make an initial independent vote; the instructor then cued students to enter into discussion in their small groups, reminding them to use reasons to back up their preferred answers; students revoted, and at the conclusion of the voting, the instructor usually called for volunteers to explain their reasoning before showing the histogram of votes. These cues closely resemble the "reasoning-centered" approach used in our previous study of advanced students (Knight *et al.*, 2013). A slight modification to this procedure was used in about one-third of the discussions: the instructor posed the question and asked students to think about it on their own but did not record an official individual vote before allowing students to move on to discussion. For these questions, students only submitted a postdiscussion vote.

Study Participants

All students in the course were asked to self-select into groups of three to four students for in-class discussion and to sit with their groups in each class period. Out of the 94 students in the lecture section studied, 23 volunteered to participate in this study (six groups). This method of participant selection was chosen for several reasons: volunteer students agreed to remain in these groups and sit in the same place in each class period, and they were willing to volunteer for the duration of the study, making it feasible to find the volunteer students in class and obtain reliable recordings. In addition, multiple recordings of several groups of students, rather than single recordings of a larger number of groups of students, allowed us to control for clicker-question variation and group-specific discussion variation in our analyses, since we collected multiple discussions for each group and multiple discussions of each clicker question from different groups.

Role and Participation of LAs

The LAs' primary role in this course was to lead problem-solving sessions outside class, with a secondary role of attending class and interacting with students during clicker discussions. LAs were either currently enrolled in or had recently completed an LA pedagogy course; they also met with the course instructors once per week to review content. For the purposes of this study, the LAs were asked to apply the training from their pedagogy course to facilitate in-class clicker discussions but were not given any additional explicit instructions. LAs were assigned to sit with a specific study volunteer group during one class period and with a nonvolunteer group during the next class period; in this way, study volunteers interacted with an LA during about half of the class periods. Because only three LAs could regularly attend this section, only three of the six volunteer groups had recurring audio recordings with and without an LA. The other three groups did not encounter an LA regularly during the period of this study, although it is possible that they may have interacted with LAs outside the class periods we recorded or in ways we could not capture with our audio devices.

Table 1. Demographics of students' class rank[a]

	n	% Female	Class rank	GPA
Nonvolunteers	71	46	1.7 (0.9)	2.8 (0.7)
Volunteers	23	52	1.8 (0.9)	3.2 (0.8)[b]

[a]Freshman = 1; sophomore = 2, etc. The volunteers are no different from the rest of the students in gender or class rank ($p < 0.05$, Mann-Whitney *U*-test).

[b]Volunteer GPA is significantly higher than nonvolunteers ($p < 0.05$, *t* test).

Data Collection

Each of the volunteer students wore a wireless microphone (lavalier style) during eight class periods (weeks 8–10 of a 15-wk semester). Demographic information provided by the registrar's office established that the volunteer students had significantly higher incoming GPAs than the section average but were otherwise representative of their section with respect to gender distribution and year in school (Table 1).

We used a Nady receiver and a digital audio recorder (Zoom Corporation) to combine wirelessly transmitted audio from each volunteer group of students during their discussions of clicker questions. The audio recordings were transcribed into an Excel spreadsheet, paired with the clicker question the students had discussed, and given a unique transcript number. Each speaker was given a number within a transcript to facilitate tracking student interactions and to tally the number of speakers per discussion. However, individual speaker identifications within a group could not be reliably preserved from discussion to discussion, precluding us from identifying and following an individual's specific contribution over time.

The time given by the instructor for each discussion and time spent in on-task discussion (Discussion Length) was noted for each recording, and Percent Productivity was calculated from these two numbers (discussion length/time given). Small deviations from the task, such as one off-topic statement, were ignored, but if students were off-topic for 30 s or more, that time was subtracted. Transcripts were then coded for features of discussion (described in *Data Analysis*), and these features were summed across discussions.

The data set for this study includes 140 discussions by six groups of students on 28 clicker questions. Owing to student absences and occasional problems with recording equipment, no single group of students was recorded discussing all 28 questions. The number of questions discussed by each group ranged from 17 to 27, as shown in Table 2. Because students voted twice on only a subset of clicker questions, this subset of 83 discussions across all groups was used to describe the impact of discussion on performance (initial vote to revote; Table 2). To explore the effect of LA presence on student discussion, we used 65 transcripts of three groups of students (12 students total) who were audio-recorded in both the presence and absence of LAs. The remaining

Table 2. Distribution of recorded discussions among different volunteer groups

	Number of discussions recorded	Number of discussions with initial and revote
Groups who interacted with an LA		
1	24	16
2	24	15
3	17	5
Total	65	36
Groups who did not interact with an LA		
4	27	17
5	21	13
6	27	17
Total	75	47
Total for both groups	140	83

75 transcripts from three groups who never interacted with LAs (11 students total) were used as a comparison group.

Finally, each clicker question was rated by two people not associated with the study as either higher-order or lower-order cognitive level, using Bloom's taxonomy (Anderson and Krathwohl, 2001; Crowe *et al.*, 2008). Raters agreed on 73% of Bloom's ratings and adjudicated all differences. Sixteen questions were rated as higher order and 12 as lower order.

Data Analysis

Discussions were coded in two ways: global, and line by line. In global coding, the transcript as a whole was characterized by the presence of each discussion characteristic. In this phase of coding, the unit of analysis was the group's whole discussion of one clicker question rather than individual student statements. LA statements were not counted as contributing to these global codes (Table 3). However, during the second phase of line-by-line coding, we characterized individual statements made by LAs and the responses made by students to each LA statement. This coding is described in more detail below in *Student Responses to Different Types of LA Statements.*

All discussion codes were developed using an iterative process, building upon our experience coding advanced student discussions (Knight *et al.*, 2013) and using a system based on Toulmin's characteristics of argumentation (Toulmin, 1958). We read through many student exchanges, discussed the interactions students engaged in, and settled on categories that were descriptive and potentially interesting. We chose to follow two global codes described previously in Knight *et al.* (2013): Exchange of Quality Reasoning and Conflicting Reasoning (which we modified into Reasoning about Multiple Answers) (Table 3). Exchange of Quality Reasoning characterizes reasoning and the use of "warrants" (Toulmin, 1958) as well as whether an exchange of such reasoning is occurring. Warrants are complete reasoning statements, in which a student provides a reason for his or her answer and connects this reason logically to data or factual information. Students can also articulate a less complete form of reasoning in which they may suggest

Table 3. Description of global codes[a]

Global code	Definition/Characteristics	Examples
Exchange of Quality Reasoning (0–3)		
0	No reason provided	"What did you vote?" "A."
1	One person provides reason(s)	"I think it's because of transcription being different." "Yeah."
2	Two or more people provide simple reason(s)	"I think it's because transcription is different in eukaryotes and prokaryotes." "Yeah, and because of the sigma factor …"
3	Two or more people provide reasons supported by evidence and a logical connection (warrants)	"I think it's because … there's no nucleus in bacteria, so that would be a difference between eukaryotes and bacteria." "Yes, there's no need to transport the transcript out of the cytoplasm since the enzyme for making the mRNA transcript is right there."
Reasoning about Multiple Answers	More than one answer is considered, using reasoning	"It doesn't have anything to do with the membrane [answer C] because …" "But I think the concentration [answer A] does matter because …"
Hedging a Reason	Signaling uncertainty in one's own reasoning	"I don't know, really, but it could be because …" "I think it works this way but I'm totally guessing"
Analogy or Example	Using an analogy or an example to help explain a reason	"It's like spraying perfume in a room."
Student–Student Questioning		
Requesting Information	Asking for votes or basic information, like definitions	"What did you vote?" "What does that mean?"
Requesting Reasoning	Asking to share an explanation	"Why did you say that?" "Why were you thinking that?"
Requesting Feedback	Asking for confirmation of own reasoning	"It takes energy to break bonds, right?"

[a]Each discussion was given a 0/1 (absence/presence) for each code, except Exchange of Quality Reasoning, as shown.

an idea or use a "because"-type statement, typically providing partial evidence supporting an idea but lacking a logical connection to their claim. We were also particularly interested in the kinds of questions that students ask one another as they discuss their ideas; the question codes we ultimately developed were based on "talk moves" used by effective K–12 teachers to prompt student discussion (Michaels and O'Connor, 2012). We iteratively refined the definition of each code as we practiced coding transcripts until the definition was clear and could be reliably recognized by multiple coders.

To establish interrater reliability, three raters each coded the same set of transcripts, discussing results and adjudicating differences over several training sessions. Interrater reliability was then established using intraclass correlation between these three raters on 24% of the total number of transcripts, achieving a Cronbach's alpha greater than 0.75 for each code reported. The line-by-line codes used to describe specific LA statements and the responses of students to those statements were developed using a similar process. These codes describe similar interactions to the global codes (e.g., reasoning, different kinds of questioning) but are used to describe each speaker's contribution to the discussion (described in more detail in *Student Responses to Different Types of LA Statements*). Two raters carried out the line-by-line coding of the entire set of transcripts independently with 86% agreement and came to consensus on any differences.

Regression Analyses

SPSS version 22 was used to run multiple regression analyses (linear, binary logistic, and ordinal logistic) to determine how each possible variable impacted the global outcome codes of the transcripts. Linear regressions were conducted for the continuous outcomes of Discussion Length and Percent Productivity. Binary logistic regressions were conducted for all presence/absence outcomes, and ordinal logistic regressions were conducted for Exchange of Quality Reasoning. The regression models generally used the following covariates: group ID, Bloom's level of question, number of speakers, and study day. Study day was assigned based on the actual day of class (i.e., class 8, 10, 17, etc.) to account for any differences over the recording time frame. An additional possible covariate was whether questions were discussed with only one vote following discussion or with an initial vote followed by discussion and revote. As this factor had no statistically significant effect on any outcomes and did not affect the significance of any models, it was not included in the final regression analyses. For linear regression models, the assumptions of linearity, independence of errors, homoscedasticity, unusual points, and normality of residuals were all met and produced models with p values < 0.01. For logistic regressions, there were no significant interactions between the continuous outcome variables, thus meeting the assumption of linearity for each analysis. For ordinal regressions, multicollinearity, proportional odds, goodness of fit, and model fitting were all within acceptable parameters (Field, 2009).

Human Subjects Approval

This work was reviewed by the University of Colorado Institutional Review Board, and the use of human subjects was approved (expedited, protocol #11-0630).

RESULTS

We first describe general student discussion characteristics and performance, then the comparisons of groups who had access to an LA during their discussions, and, finally, a detailed analysis of student responses to specific types of LA prompts.

Discussion Characteristics

Students generally began discussing a clicker question within a few seconds of being prompted by the instructor and stayed on task most of the time. Sometimes, students exchanged off-topic talk before they began discussion on the clicker question, paused in their discussion, or, in a few instances, became derailed in the middle of their discussion by a humorous remark or someone telling a story. In most cases, students discussed a topic until they reached a decision, at which point they might engage in social talk. The total time given by an instructor to any discussion depends on factors such as how quickly student votes were recorded, or whether the instructor had a conversation with a group of students during the voting period, and thus let the discussion continue for longer than normal. Overall, students spent an average of 1.25 min (± 0.68) in discussion, amounting to 64% of the time the instructor allotted (2 min ± 0.86). Notably, students engaged in discussions relevant to the clicker question 93% of the time.

To understand more about the nature of student discussion, we characterized each transcript, using seven global codes that reveal how students use reasoning and questioning: Exchange of Quality Reasoning, Reasoning about Multiple Answers, Hedging a Reason, using Analogy or Example, Requesting Information, Requesting Reasoning, and Requesting Feedback (each described in detail in Table 3). We calculated the frequency with which each of these global codes appeared in the data set of 140 discussions (Table 4). Overall, students used some kind of reasoning (levels 1–3) in 91% of their discussions but only exchanged warrants (level 3) 18% of the time. Many discussions (42%) contained a level 2 Exchange of Quality Reasoning, in which students exchanged reasons for their ideas but not necessarily warrants. A second measure of reasoning was Reasoning about Multiple Answers, in which reasons for more than one answer were considered by the group. This characteristic correlated with the level of Exchange of Quality Reasoning achieved in the discussion: in level 1 discussions, only 35% used Reasoning about Multiple Answers; thus, only one person used reasoning but offered or explained reasoning about multiple possible answers to the rest of the group. In level 2 discussions, 75% included reasoning about multiple answers, and in level 3 discussions, 100% included this type of exchange.

Two additional discussion characteristics were relatively infrequent: Hedging a Reason and using Analogy or Example. Hedging a Reason, in which students softened their reasoning with qualifiers such as "that's what I think, but I could be wrong," occurred in only 14% of discussions. Using an Analogy or Example, for example, likening the diffusion of ions to the diffusion of perfume, was similarly infrequent. Although these characteristics are potentially interesting, we chose not to follow them further with regression analyses.

Table 4. Frequency of global codes in student discussions

Whole discussion codes	Total frequencies (all 140 discussion)	Frequency of each behavior (%) among groups who interacted with LAs ($n = 65$ discussions)	Frequency of each behavior (%) among groups who did not interact with LAs ($n = 75$ discussions)
Exchange of Quality Reasoning			
0	9	9	9
1	31	40	23
2	42	37	47
3	18	14	21
Reasoning about Multiple Answers	61	57	65
Hedging a Reason	14	12	16
Analogy or Example	11	9	13
Questioning			
Requesting Information	66	74	60
Requesting Reasoning	29	50	22
Requesting Feedback	41	51	32

The last set of characteristics described how students used questioning in their discussions. Questioning was used in almost every discussion, and the three types of questions we coded were frequently used in combination: 94% of discussions included more than one type of question, with Requesting Information the most common, followed by Requesting Feedback, and then Requesting Reasoning (Table 4). A typical sequence of questions in a discussion often followed this pattern: one or more information requests, each followed by statements of the answers chosen (e.g., "What did you vote?" "C." "What about you?" "A."), then a direct request for reasoning ("Why did you think it was C?"), and a statement of a reasoning that sometimes included a request for feedback ("It is A because of [reasoning statement], right?"). Such an exchange often would end with students agreeing that the reason made sense, but sometimes another round of questioning would ensue if an alternative idea was proposed.

Clicker-Question Performance

During the study period, students answered 28 clicker questions. In 20 of these, students submitted both an initial vote and a revote following discussion. Altogether, students averaged 50% correct on their initial votes and 71% correct on their revotes. In comparing volunteer students with the rest of the students in the course, both volunteers and nonvolunteers improved in their average percent correct from initial to revote, and there is no significant difference in the two groups for either initial or revote ($p = 0.45$; Figure 1). In addition, we note that the average percent correct for questions on which there was no initial vote (77%) was similar to that of questions for which there was both an initial and revote.

To determine the potential factors affecting students' percent correct on revotes, we looked at the subset of discussions for which we had both initial and revotes recorded, and more than one person per group responding. For these 81 discussions of 20 questions, we could calculate the initial and revote percent by group and determine whether the students were in agreement on their initial vote (irrespective of correctness). Before discussion, students were in unanimous agreement on their initial vote only 25% of the time. Using a linear regression with initial agreement (0/1), initial percent correct by group, group ID, Bloom's level of the question, number of speakers, Discussion Length, and Ever with LA as covariates, we found only the initial percent correct significantly predicted the percent correct on the revote (beta = 0.43; $F(7, 73) = 3.52$, $p < 0.005$). In this model, ~20% of the variance was accounted for by the independent variables (adjusted $R^2 = 0.19$). We also explored whether any of the coded discussion features (such as Exchange of Quality Reasoning) could predict higher percent correct on revotes; these linear regression models had adjusted $R^2 = 0.18$–0.23, but no discussion features predicted the percent correct (except the covariate initial percent correct, described earlier).

Differences between Groups Who Regularly Interacted with LAs and Those Who Did Not

Because we recorded three groups (11 students) who never interacted with LAs during our recording sessions, and an

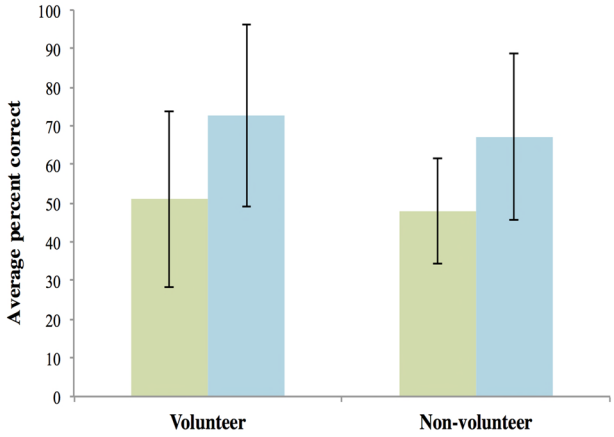

Figure 1. Average percent correct on initial (green) and revotes (blue) for the 20 clicker questions for which two votes were taken. Revote percent correct is significantly higher than the initial percent correct for both volunteers and nonvolunteers ($p < 0.01$). Volunteer and nonvolunteer measures are not different for either measure; two-way repeated analysis of variance ($p = 0.45$). Error bars show SD.

Table 5. Regression table for all 140 discussions[a]

	Exchange of Quality Reasoning	Reasoning about Multiple Answers[b]	Requesting Information[b]	Requesting Reasoning[b]	Requesting Feedback[b]	Discussion Length[c]	Percent Productivity[c]
Regression factors	Odds ratio (*p* value)	Odds ratio (*p* value)	Odds ratio (*p* value)	Odds ratio (*p* value)	Odds ratio (*p* value)	Beta (*p* value)	Beta (*p* value)
Ever with LA	0.52 (0.27)	0.67 (0.29)	1.78 (0.15)	**2.53 (0.03)***	1.97 (0.07)	0.11 (0.17)	**0.20 (0.01)***
Group ID	(0.39)[d]	1.03 (0.68)	0.94 (0.43)	1.13 (0.16)	1.15 (0.07)	–0.08 (0.34)	–0.05 (0.54)
Bloom's level (high)	1.08 (0.82)	0.49 (0.07)	1.07 (0.87)	0.94 (0.88)	0.64 (0.24)	0.03 (0.70)	–0.14 (0.06)
Number of speakers	**1.67 (0.02)***	**1.77 (0.02)***	**2.09 (0.00)***	**2.59 (0.00)***	**1.67 (0.03)***	**0.39 (0.00)***	**0.29 (0.00)***
Study day	**0.94 (0.00)***	**0.93 (0.00)***	**1.05 (0.02)***	**0.95 (0.04)***	0.99 (0.58)	**–0.23 (0.01)***	**0.31 (0.00)***

[a]Odds ratios and *p* values are shown for each factor's impact on the coded discussion features shown, using ordinal, logistic, or linear regressions. Asterisks (*) and bold type indicate significant *p* values.

[b]For logistic regressions, odds ratios < 1 indicate an inverse relationship.

[c]For linear regressions, a negative beta value indicates an inverse relationship. The linear regression models were both significant ($F(5, 134) = 10.4$, $p < 0.001$), with adjusted R^2 values of 0.14 for Discussion Length and 0.25 for Percent Productivity.

[d]For independent variables with more than two groups in an ordinal regression, an odds ratio cannot be calculated for the variable, only for each individual group. Instead, the impact of the variable can be represented by the Wald statistic: in this case, Wald $\chi^2(5) = 4.08$. The *p* value for this analysis suggests that group ID does not have a statistically significant effect on the prediction of use of higher level of reasoning.

additional three (12 students) who regularly had interactions with LAs, we compared the discussions of these two sets of students to determine whether there were any inherent differences in the two groups. From the frequency data, we noted that groups regularly interacting with LAs ("Ever with LA" condition) appeared to use less reasoning and more questioning in discussion compared with those for whom no discussions involved an LA (Table 4). To determine whether these differences were predicted by the presence of LAs with these groups, we performed an ordinal logistic regression with Exchange of Quality Reasoning as an outcome, binary logistic regressions with the presence/absence outcomes, and linear regressions with Discussion Length and Percent Productivity. The Ever with LA condition significantly influenced two outcomes: Requesting Reasoning and Percent Productivity. Groups who regularly worked with an LA were 2.5 times more likely to use Requesting Reasoning, all else being equal, and had a higher Percent Productivity (71%) than those who did not interact with LAs (59%; Table 5). Two other covariates also impacted discussion characteristics: the number of students involved in the discussion and the study day (Table 5).

Impact of LAs on Discussion

We focused the rest of our analyses on the subset of 65 discussions from the three groups who interacted with an LA. As an LA was available to any group during only about half of the recorded class sessions, this data set allowed us to estimate the impact of LAs on student discussion while controlling for group.

LA presence significantly influenced four discussion outcomes: Requesting Information, Requesting Feedback, Discussion Length, and Percent Productivity. In the presence of LAs, student discussions were 5.6 times *less* likely to use Requesting Information and 3.9 times *more* likely to use Requesting Feedback than in the absence of LAs, all else being equal (Table 6). These discussions were also significantly longer (1.58 min ± 0.6) than those without LA interaction (0.96 min ± 0.52 and more productive (78%) than those in the absence of LAs (63%), all else being equal (Table 6). In addition, more discussions involved an exchange of warrants (level 3: 20% vs. 10%), but this difference was not significant.

Some additional covariates were also significant predictors of certain discussion features. Students were significantly more likely to include Reasons for Multiple Answers when the clicker question was lower-order Bloom's rather than higher-order Bloom's. The number of speakers was a significant predictor of higher use of Requesting Information and Requesting Reasoning, and Group ID significantly predicted higher use of Exchange of Quality Reasoning and Requesting Feedback (all shown in Table 6). Finally, we also found that study day significantly affected several outcomes, sometimes positively and sometimes negatively (Table 6).

Student Responses to Different Types of LA Statements

The above analyses indicate that LAs had few significant effects on global features of student discussions. However, if LAs were using a variety of prompting statements when interacting with students, these different prompts could result in different student discussion characteristics. More detailed analysis of LA–student interactions showed this to be the case. Two different discussions of the same clicker question are shown in Figure 2 to illustrate the different responses of students to different LA prompts. In example A, the LA's initial interaction with students is to provide a reasoning statement, which does not prompt student interaction, while in example B, the LA uses several question prompts to draw out student ideas, resulting in students exchanging ideas.

To better understand how LA prompting statements impacted student responses, we characterized both using the line-by-line codes shown in Table 7. LA statements were coded into five categories: Prompting Questions (asking students which answer they selected or asking them to consider some additional piece of information), Requesting Reasoning (asking questions to elicit reasoning), Background Statement (a statement of factual information but not a reason), Providing Reasoning (explaining a reason for an answer),

Table 6. Regression table for conversations with an LA ($n = 33$) and without an LA ($n = 32$) for the three groups who interacted with LAs[a]

	Exchange of Quality Reasoning	Reasoning about Multiple Answers[b]	Requesting Information[b]	Requesting Reasoning[b]	Requesting Feedback[b]	Discussion Length[c]	Percent Productivity[c]
Regression factors	Odds ratio (p value)	Odds ratio (p value)	Odds ratio (p value)	Odds ratio (p value)	Odds ratio (p value)	Beta (p value)	Beta (p value)
LA present	2.76 (0.08)	2.37 (0.18)	**0.19 (0.04)***	0.79 (0.69)	**3.19 (0.05)***	**0.46 (0.00)***	**0.24 (0.05)***
Group ID	**(0.05)*,[d]**	1.18 (0.11)	0.86 (0.19)	1.13 (0.22)	**1.24 (0.03)***	–0.03 (0.83)	–0.07 (0.55)
Bloom's level (high)	0.72 (0.49)	**0.18 (0.01)***	1.57 (0.51)	0.82 (0.74)	0.74 (0.60)	0.05 (0.65)	–0.01 (0.35)
Number of speakers	1.35 (0.36)	0.96 (0.93)	**3.94 (0.01)***	**2.72 (0.02)***	0.72 (0.39)	0.14 (0.26)	0.16 (0.22)
Study day	0.95 (0.08)	**0.93 (0.04)***	1.07 (0.08)	**0.93 (0.05)***	1.02 (0.74)	–0.21 (0.06)	**0.39 (0.001)***

[a]Odds ratios and p values are shown for each of the factor's effects on the characteristics of student discussion using ordinal, binary logistic, or linear regressions. Asterisks (*) and bold type indicate significant p values.

[b]For logistic regressions, odds ratios < 1 indicate an inverse relationship. Thus, in each of these cases, the odds ratios can be inverted to better describe the outcome: for example, lower-level Bloom's questions are 5.6 times more likely than higher-level questions to generate reasoning about multiple answers, all else being equal; and the absence of an LA is 5.3 times more likely to generate requests for information, all else being equal.

[c]For linear regressions, a negative beta value indicates an inverse relationship. The linear regression models were both significant ($F(5, 59) = 4.6$, $p = 0.001$), with adjusted R^2 values of 0.25 for Discussion Length and 0.22 for Percent Productivity.

[d]For independent variables with more than two groups in an ordinal regression, an odds ratio cannot be calculated for the variable, only for each individual group. Instead, the impact of the variable can be represented by the Wald statistic: in this case, Wald $\chi^2(2) = 5.98$. The p value for this analysis suggests that group ID has a statistically significant effect on the prediction of use of higher level of reasoning.

or Acknowledgment (a simple statement, such as "yes"). Student responses were also coded into five similar but not identical categories. Because students did not directly ask LAs to explain their reasoning, we combined the two questioning categories into a single category (Asking Questions), and an End of Discussion category was added to track circumstances in which an LA statement was followed by no substantive student contributions. Otherwise, the categories of Using/Providing Reasoning, Background Statements, and Acknowledgment were the same as for the LA statements.

LAs engaged with students on average three times per discussion, for a total of 110 LA statements across the 33 discussions. The most common type of LA statement was a Prompting Question, asking students to share what they voted or to use information they had not yet considered (Table 8). This prompt was typically used by LAs to initiate discussion with the students: in one-third of the discussions, the LAs used a sequence of at least two questions before providing reasoning or background. LAs also frequently used background statements to provide additional information in response to student questions or to prompt additional thinking. In another third of the discussions, the LAs asked a combination of one or more questions and other statements, and then offered their own reasoning. Although LAs were

Clicker question: Given what you know about the chemical properties of the lipid bilayer, which of the following proteins is UNLIKELY to be found associated with the membrane?

a. A protein with one or more stretches of ~20-30 consecutive hydrophobic amino acids
b. A globular protein with nonpolar amino acid chains folded into the interior of the protein and hydrophilic side chains exposed to water
c. A protein consisting of ONLY hydrophilic amino acids (no hydrophobic amino acids)
d. A protein with 20-30 amino acid stretch of alternating polar and nonpolar amino acids
e. None of the above.

LA provides reasoning

1: What did you think it was?

2: I'm saying C, a protein consisting of only hydrophilic...

LA: So think about…. associating with the membrane can mean within the membrane and it can also be outside the membrane, right? So if it's associating with the membrane it can be a transmembrane protein and it can also be a peripheral membrane protein, 'cause it's still interacting with the membrane.

2: Oh, so I am saying E, it's none of them.

LA: Yeah, I like E.

LA asks prompting questions

1: So its not A, because that can't exist, B seems like it can exist… I… don't know, well I mean it's not the answer. Let's see, hydrophilic interactions… you can't have C it would just be outside of the membrane, I think.

2: I don't think the answer is D.

LA: You want to think about what proteins do when they associate with the membrane, so… what does D describe?

2: Ummm… A trans protein, well when it goes in and out of the cell…I don't remember what that's called.

LA: Yeah and what's…

1: Well, D is not long enough to go in and out several times, cause it's only one stretch of 20-30 and that's like one membrane thick, right? So, E.

Figure 2. Two examples of different groups of students interacting with an LA in discussing the same clicker question. In the first case, the LA provides reasoning, explaining the answer to the students. In the second case, the LA uses prompting questions, and the students respond by exchanging reasoning for their answers.

Table 7. Line-by-line codes used to describe individual LA statements and the responses of students

LA statement	Definition /characteristics	Examples
Prompting Question	Request for a student's answer; request for information	"What did you answer?" "What do you think the differences are between those two?"
Requesting Reasoning	Request for sharing an explanation or otherwise providing support for a claim	"Why did you pick C and not D?"
Providing Reasoning	Provides explanatory statement of backing, evidence, or justification	"Your genetic code is consistent throughout all organisms … all that really matters is the gene itself."
Background Statement	Shares basic information	"The genetic code just says UAC codes for tyrosine …"
Acknowledgment	Acknowledgment of a statement	"Yes, that sounds right."
Student response		
Asking Question	Request for basic information, an explanation, or confirmation of reasoning	"What does that word mean?" "Why do you think that?" "It takes energy to break bonds, right?"
Using Reasoning	Provides explanatory statement of backing, evidence, or justification	See Exchange of Quality Reasoning for examples
Background Statement	Student sharing basic information or states support of an answer	"I think the sigma factor is for eukaryotes."
Acknowledgment/Claim	Acknowledgment of a statement or statement of a clicker vote choice	"Yes." "I picked C."
End of Discussion	When statement of affirmation ends the discussion	"Okay, that makes sense."

encouraged in their general pedagogical training to question and prompt rather than explain the reason for an answer, LAs contributed reasoning statements in more than half of the discussions. However, in only one discussion did an LA begin by providing reasoning. Finally, the least common LA interaction beside simple acknowledgment was directly requesting students' reasoning.

Each LA statement produced an average of 1.4 student responses, with a total of 156 student responses to the 110 LA statements. To characterize student responses to LA statements, we tracked which student responses followed each type of LA statement until an LA spoke again. For example, in response to Prompting Questions from LAs, student responses included 20 questions, 10 reasoning statements, six background statements, 17 acknowledgments, and one end of discussion. Because acknowledgment statements, although common, do not contribute substantively to the discussion, we did not follow this category further.

The frequency with which each type of LA statement was followed by each student response is shown in Figure 3.

Table 8. Frequency of use of the five types of LA statements[a]

LA statement category	Percent of 33 discussions in which LA statement was used	Percent of 110 total LA statements	Number of student responses
Prompting Question	64	30	53
Request for Reasoning	39	17	24
Using Reasoning	52	25	35
Background Statement	59	21	36
Acknowledgment	21	7	9

[a]More than one type of LA statement was used in 29 of 33 discussions. The total number of student responses to each type of LA statement is also shown.

When LAs asked a prompting question or provided a background statement, students were most likely to ask a question, while students were most likely to respond with a reasoning statement when LAs directly requested that the students provide reasoning. None of these LA behaviors was likely to be followed by the end of student discussion ($p < 0.05$ for all of the above, chi-square test). However, when LAs explained their reason for an answer, students were *least* likely to respond with their own reasoning or background statements ($p < 0.05$, chi-square test; Figure 3). Using pairwise comparisons of the responses to each type of LA statement, LA-provided reasoning was significantly more likely to result in an end of student discussion than any other LA statement. In addition, an LA requesting reasoning was significantly more likely to elicit student reasoning than any other LA statement ($p < 0.05$, Mann-Whitney *U* test; Figure 3).

DISCUSSION

Previous research, primarily in K–12 (Jimenez-Aleixandre *et al.*, 2000; Sandoval, 2003) but increasingly in undergraduate education (e.g., James and Willoughby, 2011; Aydeniz *et al.*, 2012; Kulatunga *et al.*, 2013), has shown the importance of discussion and argumentation among students as a factor in helping them learn both scientific content and the practice of science. Even when discussions do not directly improve short-term performance on test items, exchanging ideas contributes to understanding the discourse practices of science and promotes learning by giving students feedback on their internal construction of ideas and a chance to hear how others are constructing their ideas (Ford and Forman, 2006; Next Generation Science Standards, 2013). In this paper, we have characterized how students discuss clicker questions in a large introductory biology course, with a particular focus on reasoning and questioning, and explored whether interactions with LAs affected student discussion patterns.

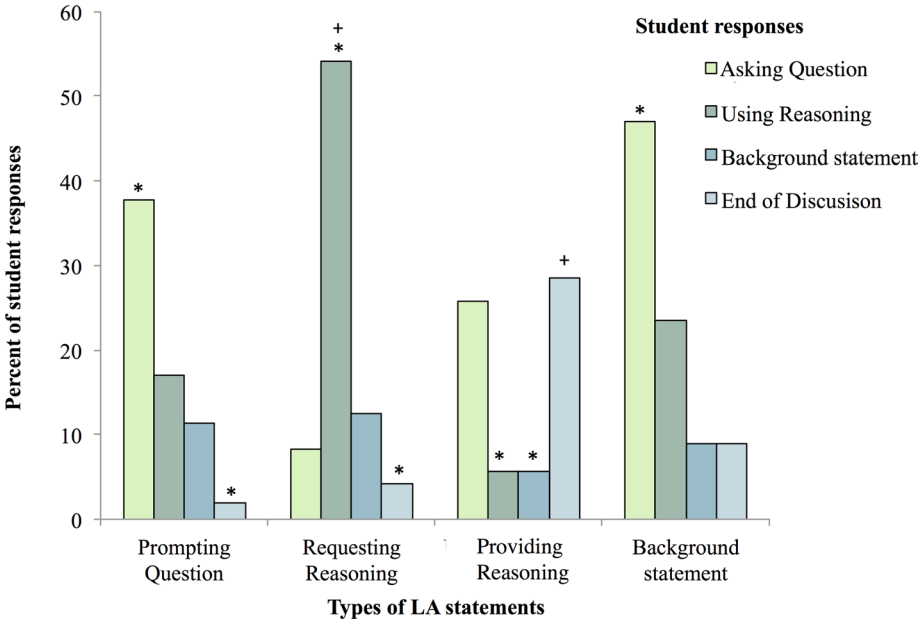

Figure 3. Each bar shows the type of student response as a percent of total student responses to each LA statement. Asterisks (*) indicate that this response differed significantly from the expected value (all being equal) to each LA statement ($p < 0.05$, chi-square test). Plus signs (+) indicate that this student response was significantly more likely in response to the indicated LA statement than in response to any other LA statement ($p < 0.04$, Mann-Whitney *U*-test).

We found that LAs can positively influence the articulation of reasoning in student discussions, especially if they use prompting questions and requests for reasoning.

Student Articulation of Reasoning

In this study, students were cued by their instructors to discuss the reasons for their answers with one another, and they did so most of the time. However, students relatively infrequently exchanged reasoning with fully articulated warrants—Exchange of Quality Reasoning level 3—claims logically connected to evidence (Table 4). Because these students did not receive explicit instruction on how to use reasoning or on how to construct an argument, this base level of exchange likely is representative of "untrained" student tendencies. Several other studies have shown (primarily with middle and high school students) that discerning what information is relevant in answering a question, using evidence to back up ideas, and justifying one's reasons are all challenging endeavors (e.g., Erduran *et al.*, 2004; McNeill and Krajcik, 2007). Students do not typically employ these behaviors without training, although they may learn to do so when they frequently work in groups (Lubben *et al.*, 2009).

One recent study of undergraduates showed that students are able to engage in higher levels of reasoning when given complex problem-solving opportunities in a format that explicitly encourages group problem solving and argumentation. Using a modified version of Toulmin's characterization of argumentation to describe instances of students coconstructing reasoning, Kulatunga *et al.* (2013) studied a small group of General Chemistry 1 students in peer-led guided inquiry sessions and found that they used coconstructed warrants 50% of the time, considerably higher than the 18% in our sample. This difference may be generated by additional instruction in how to engage in argumentation, differences in the kind of tasks the students were engaged in, or the time given for discussion. These findings support our suggestion that instructional design and practice can substantially influence student use of argumentation.

In comparing different levels of undergraduate biology students, introductory and advanced students clearly differ. Advanced students have engaged in more years of course work and have likely practiced the use of reasoning, even if they have not explicitly been taught to use argumentation. In comparing the introductory students in this study with the advanced students from our previous study, all of whom received a similar instructor cue to consider the reasons for their answers, advanced students clearly use more level 3 exchanges of quality reasoning (more than 50% of discussions; Knight *et al.*, 2013), versus 18% for the introductory students in the current study. While advanced students used predominantly level 3 (exchanges of warrants), introductory students used mostly level 2, in which they exchanged reasons that lacked either the evidence or logical connection to qualify as warrants. Introductory students also used more level 1 reasoning (one student explaining his or her reasoning to others) than did advanced students. This may reflect, in part, the more heterogeneous preparation of introductory students. Some introductory students may feel more knowledgeable or actually be more knowledgeable than others; these students may be more confident in taking charge of discussions and teaching other students. In an advanced major's class, in which all students have taken the same prerequisites and have had more experiences engaging in discussion, more students may feel confident contributing reasons for their answers or refuting others' ideas.

Other studies have reported difficulties in training students to recognize features of a good argument. Van Lacum *et al.* (2014) trained students to follow seven different argumentation characteristics in reading original science literature and tested students both before and after instruction to measure the impact of such training. Students improved in their ability to identify the motive, objective, main conclusion, and implications in the papers they read, but did not improve in their ability to identify supports (warrants) and

counterarguments within the writing. Although students were identifying written versions of argumentation practices rather than engaging in exchanges of ideas, the study highlights that it is challenging for students to learn how to identify and use the building blocks of a high-quality argument.

Discussion Time

The students in our study spent a relatively short time discussing their ideas, ~1 min on average, unless LAs were present, in which case they spent ~1.5 min. These times are similar to the length of time reported in other studies of introductory students answering clicker questions (e.g., introductory physics; Miller *et al.*, 2014). Although we only tracked performance on clicker questions by group, rather than by individual student performance, we did not find any correlation between the time taken in discussion and the correctness of the group in its answer. In contrast, Miller *et al.* (2014) found that individual students who answered clicker questions incorrectly took longer to record their answer, both before and after peer discussion. However, these authors did not measure the length of actual peer discussion, only each individual's time to response after clicker polling was opened.

Introductory biology students could have relatively short discussions for many possible reasons. Students may not be aware of their lack of understanding (lack of metacognition) and thus may not engage in lengthy discussion of their ideas. They may not feel comfortable challenging incorrect peer ideas or raising alternative ideas. In addition, they may not discuss their answers for a longer time, because they are not concerned about whether they get the question correct or incorrect (in this study, clicker points were for participation only). Assured of participation credit, students might choose to hear the class discussion or instructor explanation of the question rather than engaging with their peers. On the other hand, previous research has shown that awarding points for correctness tends to lead to consensus building around a dominant speaker's ideas, rather than coconstructed reasoning (James, 2006), which could shorten discussion even more.

The time taken by introductory students to discuss their ideas was considerably shorter than previously found among advanced biology students, who took 2.5 min on average, and sometimes more than 4 min, to discuss their ideas (Knight *et al.*, 2013). The questions that the advanced students answered were generally more cognitively demanding: only 20% of the questions were lower order, while 43% of the questions asked of introductory students were lower order. Thus, it is possible that, because introductory students were generally answering less cognitively demanding questions, they needed less time in discussion. An equally plausible conjecture is that advanced students spend more time because they are more engaged in the process of discussion, more engaged in the material, and/or simply have more practice and comfort engaging in the process of argumentation. Finally, instructors may set a norm for the classroom by wrapping up discussions after 1–2 min in the first few class periods, leading students to expect this relatively short amount of time.

Performance

Performance on clicker questions does not appear to be impacted directly by any features of student groups (such as number of speakers, group ID, or presence of an LA) or by whether the group was in agreement before beginning the discussion. Only the initial percent correct (by group) predicted higher revote percent correct. In addition, no other coded features of the discussion predicted the performance outcome on a question. For example, the level of reasoning achieved in a discussion did not have a reproducible affect on clicker-question performance, even though one might have predicted more exploration of reasoning would result in more correct answers. The lack of connection between discussion features and correctness of a group's answers has been found by us with advanced students (Knight *et al.*, 2013) and by others with younger students (e.g., Sampson and Clark, 2008), suggesting it is not a unique feature of any particular level of student. Sampson and Clark (2008) found that students who worked collaboratively to generate arguments performed better later on transfer tasks than those who worked individually, but their arguments did not necessarily reflect deeper understanding or better lines of evidence compared with individual student work. The authors point out that students sometimes ignored ideas proposed by classmates or failed to engage in critical discussion of proposed ideas, thus not necessarily constructing a complete understanding of the question at hand. However, their improved performance on later individual transfer tasks indicates that their argumentation in some way enhanced their learning.

Therefore, lack of connection between rich student discussion and a single measure of immediate question correctness does not imply that the discussion is not helping students learn. Because the clicker-question responses are not graded, students may feel free to discuss and vote without concern for whether they are correct. Many other factors likely influence whether students fully process the ideas discussed. Ultimately, determining whether discussions involving high-quality reasoning and exchange of questions result in improved undergraduate student performance will likely require using multiple different and longer-term measures of assessment in addition to the immediate in-class diagnostic.

Cognitive Difficulty

As previously shown (Knight *et al.*, 2013), the Bloom's level of a question does not necessarily predict the characteristics of student discussion. However, when looking at the subset of groups who engaged in discussion with LAs about half the time, if the question was a *lower*-order Bloom's level, students were significantly more likely to articulate reasoning for multiple answers. This is somewhat surprising, in that we would have assumed that the more cognitively demanding questions would generate more discussion of different answers. However, introductory students may find it easier to discuss a question that experts consider fact based, or even easy, because they may feel confident in helping one another clarify factual information required to answer the question. Thus, they may engage in an exchange of information about the likelihood of each answer as they try to figure out what each answer means. Cognitively challenging questions may also be more difficult for groups to initiate or sustain a discussion around, as the argument needed to support an answer would be more complex. The implication of this finding is that in courses emphasizing cognitively challenging questions, students may require more explicit instruction on how to engage in argumentation.

Group Size

In the discussions we recorded, an average of 3.3 students participated, and the number of speakers across all discussions ranged from one to four speakers, with an occasional discussion involving five. When considering all discussions, a higher number of speakers involved in discussion was positively correlated with all global discussion characteristics (Table 5). When considering a subset of the discussions (only the groups who had a chance to interact with an LA; Table 6), the number of speakers predicted only higher use of requesting information and requesting reasoning but did not have a significant impact on any other characteristics. These patterns were found in an auditorium-style lecture hall, indicating that seating arrangement is not necessarily a limitation to productive discussion among larger groups. This finding aligns with previous studies suggesting that groups should be between three and five to maximize student participation and exchange of ideas (Beichner and Saul, 2003). With more speakers in a conversation, it is more likely that different ideas will be expressed, more questions asked, and potentially higher level exchanges of reasoning. Groups of more than five may be too big to be productive, but we were not able to measure discussions among larger groups. This finding does suggest that one course dynamic component to consider should be the size of groups, with four to five students potentially being better than two or three for stimulating exchanges of ideas.

Study Day

We did not necessarily expect student discussions to improve in quality of reasoning or exhibit more of the other coded characteristics over the short recording period (2.5 wk). However, some topics are much more challenging for students than others, and this could be reflected in the students' discussions of clicker questions asked on different topics. Thus, we included study day as a factor in the regression analyses and found that this factor was a significant predictor of several discussion characteristics, sometimes inversely. In none of these cases were the odds ratios or betas very high, indicating a limited effect of this variable (Tables 5 and 6). We suggest that the characteristics of student discussions are likely to vary substantially from day to day due to the nature and perceived difficulty of the content being discussed in that class period. Because we can hold this factor constant in the regression analyses while exploring whether other factors impact each discussion characteristic, it does not confound our other findings. However, this finding does support our initial rationale for collecting multiple data points for individual groups of students, as their discussions are different on the different topics. It is also possible that student discussions change substantially over the course of an entire semester: students may better learn how to better interact with one another and may take more or less of an interest in their discussions. To measure this potential, one would need to examine the difference between discussions students have early in a semester (weeks 2–3) to the discussions the same students have much later in a semester (weeks 10–12).

Impact of LAs

At the beginning of our study, we predicted that LA interactions would enhance global reasoning-related discussion features such as Exchange of Quality Reasoning. Instead, we found that LA presence increased student use of reasoning, but not significantly (Table 6). On the other hand, LA presence did impact student use of questioning, shifting them away from using information questions such as "What did you vote?" to questions that were reasoning related (i.e., reasoning statements followed by a feedback request). This increased use of feedback questions is not surprising, as the simple presence of an LA who has previously taken the course is likely to prompt questions from the students. Students also spent significantly more time in discussion when an LA was present; this difference is specific to when the LA is actually interacting with students (Table 6) rather than being a function of an individual group's dynamics (Table 5). Thus, having an LA engage in discussion with students has the potential to help them discuss the material more thoroughly. However, it is also clear that LAs interact with students in heterogeneous ways, likely accounting for the lack of a significant effect on the Exchange of Quality Reasoning in discussion.

The LAs involved in this course are generally between one semester to 2 yr more advanced than the students enrolled in the course. Thus, they are by no means content experts or experts at teaching, which many studies suggest requires extensive training (e.g., Spillane, 1999; Windshitl, 2004). Perhaps not surprisingly, then, despite pedagogical instruction, LAs do not perfectly execute their interactions with students. In our sample, LAs explained answers to students, using reasoning, in 50% of their discussions. LAs likely think they are helping students build understanding by explaining an answer, when, in fact, the data indicate that providing reasoning often ended the students' attempts to grapple with the material. Importantly, however, LAs can have a positive effect on student interactions when they use practices that have been demonstrated to draw students into deeper cognitive processing. When LAs used prompting questions, either to encourage student thinking or to more explicitly ask for their reasoning, students readily engaged in these behaviors. A recent study on chemistry peer tutors and their interactions with students during process-oriented guided inquiry learning activities (Kulatunga and Lewis, 2013) showed markedly similar outcomes. In their study, the authors identified a suite of interactions that peer tutors used with students, most of which overlap with our classifications. These authors found that "direct teaching" (our Providing Reasoning category) was much less successful at getting students to generate warrants than "probing and clarifying" interactions (our Prompting Question and Requesting Reasoning categories). We strongly support their conclusion that, when peer tutors combine questioning and mediating behaviors, they are creating an instructional scaffolding that helps students produce high-quality reasoning and deeper understanding and that serves to generate more student questions. All of these interactions are likely to promote the use of reasoning, and help students learn from their in-class discussions.

Caveats

In this study, we collected data from discussions among a relatively small group of students in a single course. It is possible that other students in different courses or even other students in the same context might behave differently. However, the large number of discussions analyzed, and the

relative similarity of the data recorded from these students to data from previously studied advanced students (Knight *et al.*, 2013) and other groups of introductory biology students in different courses and years (unpublished data), indicates that these students are likely to be representative of other students in other courses under similar conditions. The strength of this study is in the *number* of discussions analyzed, as we can generalize the discussion trends and characteristic interactions for the students who were observed. The coding scheme we have developed and the variables that impact discussion can be used by others to document their own students' interactions and/or test for additional factors we did not measure.

Instructional Implications

While components of argumentation may be part of student thinking processes, they are not often articulated fully during in-class discussions. On the other hand, students are willing and able to engage in argumentation if given the tools and time to do so. If instructors agree there are benefits to engaging students in reasoning during class, they will want to make this connection explicit in their learning objectives and in establishing classroom norms. To increase the use of reasoning in the classroom, instructors can use more specific cues to trigger student articulation of reasoning and can train students in how to construct an argument, taking class time to show students the utility of this practice.

Relatively small changes to cues and practices have the potential to make a difference. For example, instructors could rotate a variety of meaningful cues before each clicker question or longer group discussion that clearly state how the students should interact (e.g., "Explain your reasoning for your answer to your neighbor, and ask for their reason even if you agree on the answer," or "With your neighbor, explain your reasons, and determine whether you have all the information you need to be confident of your answer," or "Find someone who chose a different answer and discuss your reasons, even if you have to ask other groups," or "Don't be hesitant to challenge someone's reason in a constructive way"). Subtle changes to what students are asked to do will keep the cues from sounding repetitive and yet continue encouraging students to practice articulating their reasoning.

In addition, seeing what scientists consider to be adequate evidence for drawing conclusions and how they convince one another of the mechanism behind a particular concept are likely essential for helping students understand how to construct their own arguments. This could be achieved in a classroom setting by showing examples of scientists constructing arguments, and then conducting a class-wide discussion of a similar question and its answers using an argumentation framework (such as Toulmin's), guided by the instructor. This type of modeling should be repeated several times during a course, so the analysis and practice of reasoning becomes a classroom norm.

Perhaps the most important implication of this study is that instructors must not rely solely on general pedagogical training of peer coaches to promote the types of interactions we expect such coaches to encourage. Using prompting or leading questions along with direct requests for reasoning may not be easy for peer coaches to implement in real class settings, even if they have learned about these practices in their training. Showing peer coaches data demonstrating that their impact can be positive or negative and that student behaviors are predicted by how they choose to engage with students may remind them to monitor their interactions. In addition, peer coaches would undoubtedly benefit from more hands-on training, with particular attention paid to practicing how to scaffold learning with their students (Kulatunga and Lewis, 2013) and using role-playing strategies to practice encouraging exchange of reasoning. The recommendations made above for instructor–student interactions can be applied to better training peer coaches as well; after all, these coaches are still students and likely need more practice in learning how to use reasoning themselves before they can help others do so. Finally, instructors and peer coaches alike would benefit from working together to explore ways to generate a classroom environment that focuses on deep understanding of content through the explicit use of argumentation and reasoning.

ACKNOWLEDGMENTS

This work was carried out with support from the National Science Foundation (DUE 1140789). We are grateful to the members of our grant advisory board: Noah Finkelstein, Derek Briggs, Valerie Otero, Melissa Dancy, and Laurel Hartley and our external evaluator, Sam McKagan, for excellent suggestions and assistance during this project. We are also indebted to Adam Bohr and Roddy Theobald for help with statistical analysis and to the Discipline Based Education Research group of the University of Colorado for ongoing feedback. For help with data collection and transcription, thanks to undergraduate assistants Amedee Martella, Alex Merritt, Nick Myers, Erika Lai, Francis Li, and Sarah Zimmermann. This work could not have been completed without the gracious support of instructors Jennifer Martin and Nancy Guild, and the willingness of University of Colorado students to participate in this research.

REFERENCES

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC.

Anderson LW, Krathwohl DR (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, New York: Longman.

Asterhan CSC (2013). Epistemic and interpersonal dimensions of peer argumentation: conceptualization and quantitative assessment. In: Affective Learning Together, ed. M Baker, J Andriessen, and S Jarvela, Advances in Learning and Instruction, New York: Routledge, 251–272.

Asterhan CSC, Schwarz BB (2009). Argumentation and explanation in conceptual change: indications from protocol analyses of peer-to-peer dialog. Cognitive Sci *33*, 374–400.

Aydeniz M, Pabuccu A, Cetin PS, Kaya E (2012). Argumentation and students' conceptual understanding of properties and interactions of gases. Int J Sci Math Educ *10*, 1303–1324.

Bao L, Cai T, Koenig K, Fang K, Han J, Wang J, Liu Q, Ding L, Cui L, Luo Y, *et al.* (2009). Learning and scientific reasoning. Science *323*, 586–587.

Beichner RJ, Saul JM (2003). Introduction to the SCALE-UP (Student-Centered Activities for Large Enrollment Undergraduate Programs) Project. Proceedings of the International School of Physics "Enrico Fermi," Varenna, Italy. www.ncsu.edu/per/scaleup.html.

Bricker LA, Bell P (2008). Conceptualizations of argumentation from science studies and the learning sciences and their implications for the practices of science education. Sci Educ *92*, 473–498.

Coleman EB (1998). Using explanatory knowledge during collaborative problem solving in science. J Learn Sci *7*, 387–427.

Crowe A, Dirks C, Wenderoth MP (2008). Biology in Bloom: implementing Bloom's taxonomy to enhance student learning in biology. CBE Life Sci Educ *7*, 368–381.

Dancy M, Henderson C (2010). Pedagogical practices and instructional change of physics faculty. Am J Phys *78*, 1056.

Duschl R (2008). Science education in three-part harmony: balancing conceptual, epistemic, and social learning goals. Rev Res Educ *32*, 268–291.

Erduran S, Simon S, Osborne J (2004). TAPping into argumentation: developments in the application of Toulmin's argument pattern for studying science discourse. Sci Educ *88*, 915–933.

Field A (2009). Discovering Statistics Using SPSS, London: Sage.

Ford MJ, Forman EA (2006). Chapter 1: redefining disciplinary learning in classroom contexts. Rev Res Educ *30*, 1–32.

Furtak EM, Alonzo AC (2010). The role of content in inquiry-based elementary science lessons: an analysis of teacher beliefs and enactment. Res Sci Educ *40*, 425–449.

Handelsman J, Ebert-May D, Beichner R, Bruns P, Chang A, DeHaan R, Gentile J, Lauffer S, Stewart J, Tilghman SM, Wood WB (2004). Policy forum: scientific teaching. Science *304*, 521–522.

Henderson C, Dancy M (2008). Physics faculty and educational researchers: divergent expectations as barriers to the diffusion of innovations. Am J Phys *71*, 79–91.

James MC (2006). The effect of grading incentive on student discourse in peer instruction. Am J Phys *74*, 689–691.

James MC, Willoughby S (2011). Listening to student conversations during clicker questions: what you have not heard might surprise you! Am J Phys *79*, 123–131.

Jimenez-Aleixandre MP, Rodrıguez AB, Duschl RA (2000). "Doing the lesson" or "doing science": argument in high school genetics. Sci Educ *84*, 757–792.

Knight JK, Wise SB, Southard KM (2013). Understanding clicker discussions: student reasoning and the impact of instructional cues. CBE Life Sci Educ *12*, 645–654.

Koslowski B (1996). Theory and Evidence: The Development of Scientific Reasoning, Cambridge, MA: MIT Press.

Kuhn D (1993). Science as argument: implications for teaching and learning scientific thinking. Sci Educ *77*, 319–337.

Kulatunga U, Lewis JE (2013). Exploration of peer leader verbal interactions as they intervene with small groups in college general chemistry. Chem Educ Res Pract *14*, 576–588.

Kulatunga U, Moog RS, Lewis JE (2013). Argumentation and participation patterns in general chemistry peer-led sessions. J Res Sci Teach *50*, 1207–1231.

Lewis SE, Lewis JE (2008). Seeking effectiveness and equity in a large college chemistry course: an HLM investigation of peer-led guided inquiry. J Res Sci Teach *45*, 794–811.

Lubben F, Sadeck M, Scholtz Z, Braund M (2009). Gauging students' untutored ability in argumentation about experimental data: a South African case study. Int J Sci Educ *32*, 2143–2166.

Mazur E (1997). Peer Instruction: A User's Manual, Saddle River, NJ: Prentice Hall.

McNeill KL, Krajcik J (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In: Thinking with Data: The Proceedings of the 33rd Carnegie Symposium on Cognition, ed. M Lovett and P Shah, Mahwah, NJ: Erlbaum.

Michaels S, O'Connor C (2012). Talk Science Primer, Cambridge, MA: Technical Education Research Center.

Miller K, Lasry L, Lukoff B, Schell J, Mazur E (2014). Conceptual question response times in peer instruction classrooms. Phys Rev Spec Top Phys Educ Res *10*, 020113.

National Research Council (NRC) (2007). Taking Science to School: Learning and Teaching Science in Grades K–8, Washington, DC: National Academies Press.

NRC (2012). Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering, Washington, DC: National Academies Press.

Next Generation Science Standards (2013). NGSS home page. www.nextgenscience.org (retrieved 14 January 2015).

Osborne J (2010). Arguing to learn in science: the role of collaborative, critical discourse. Science *328*, 463–466.

Osborne J, Erduran S, Simon S (2004). Enhancing the quality of argumentation in school science. J Res Sci Teach *41*, 994–1020.

Otero V (2006). The Learning Assistant model for Teacher Education in Science and Technology. American Physical Society. www.aps.org/units/fed/newsletters/summer2006/otero.html (retrieved 14 October 2015).

Otero V, Pollock SJ, Finkelstein N (2010). A physics department's role in preparing future teachers: The Colorado Learning Assistant Model. Am J Phys *78*, 1218–1224.

President's Council of Advisors on Science and Technology (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics, Washington, DC: U.S. Government Office of Science and Technology.

Quitadamo IJ, Brahler CJ, Crouch GJ (2009). Peer-led team learning: a prospective method for increasing critical thinking in undergraduate science courses. Sci Educator *18*, 29–39.

Sampson V, Clark D (2008). The impact of collaboration on the outcomes of scientific argumentation. Sci Educ *93*, 448–484.

Sandoval WA (2003). Conceptual and epistemic aspects of students' scientific explanations. J Learn Sci *12*, 5–51.

Smith MK, Wood WB, Adams WK, Wieman C, Knight JK, Guild NA, Su TT (2009). Why peer discussion improves student performance on in-class concept questions. Science *323*, 122–124.

Smith MK, Wood WB, Krauter K, Knight JK (2011). Combining peer discussion with instructor explanation increases student learning from in-class concept questions. CBE Life Sci Educ *10*, 55–63.

Songer NB, Gotwals AB (2012). Guiding explanation construction by children at the entry points of learning progressions. J Res Sci Teach *49*, 131–165.

Spillane JP (1999). External reform initiatives and teachers' efforts to reconstruct their practice: the mediating role of teachers' zones of enactment. J Curric Studies *31*, 143–175.

Toulmin S (1958). The Uses of Argument, Cambridge, UK: Cambridge University Press.

Turpen C, Finkelstein ND (2009). Not all interactive engagement is the same: variations in physics professors' implementation of peer instruction. Phys Rev ST Phys Educ Res 5, 020101.

Turpen C, Finkelstein ND (2010). The construction of different classroom norms during peer instruction: students perceive differences. Phys Rev ST Phys Educ Res 6, 020123.

Van Lacum EB, Ossevoort MA, Goedhart MJ (2014). A teaching strategy with a focus on argumentation to improve undergraduate students' ability to read research articles. CBE Life Sci Educ *13*, 253–264.

Windschitl M (2004). Folk theories of "inquiry": how preservice teachers reproduce the discourse and practices of the scientific method. J Res Sci Teach *41*, 481–512.

Zohar A, Nemet F (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. J Res Sci Teach *39*, 35–62.