CBE—Life Sciences Education
Vol. 12, 645–654, Winter 2013

Article

Understanding Clicker Discussions: Student Reasoning and the Impact of Instructional Cues

Jennifer K. Knight,* Sarah B. Wise,† and Katelyn M. Southard‡

*Department of Molecular, Cellular, and Developmental Biology and †Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, CO 80309; ‡Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721

Submitted May 6, 2013; Revised June 11, 2013; Accepted July 2, 2013
Monitoring Editor: Clarissa Ann Dirks

Previous research has shown that undergraduate science students learn from peer discussions of in-class clicker questions. However, the features that characterize such discussions are largely unknown, as are the instructional factors that may lead students into productive discussions. To explore these questions, we recorded and transcribed 83 discussions among groups of students discussing 34 different clicker questions in an upper-level developmental biology class. Discussion transcripts were analyzed for features such as making claims, questioning, and explaining reasoning. In addition, transcripts were categorized by the quality of reasoning students used and for performance features, such as percent correct on initial vote, percent correct on revote, and normalized learning change. We found that the majority of student discussions included exchanges of reasoning that used evidence and that many such exchanges resulted in students achieving the correct answer. Students also had discussions in which ideas were exchanged, but the correct answer not achieved. Importantly, instructor prompts that asked students to use reasoning resulted in significantly more discussions containing reasoning connected to evidence than without such prompts. Overall, these results suggest that these upper-level biology students readily employ reasoning in their discussions and are positively influenced by instructor cues.

INTRODUCTION

Many lines of research have shown that a wide variety of in-class active-learning activities designed to engage students through discussion of conceptually challenging questions lead to improved student outcomes in undergraduate biology classes (e.g., Udovic *et al.*, 2002; Kitchen *et al.*, 2003; Freeman *et al.*, 2007; Walker *et al.*, 2008). Studies of the use of clickers, remote response systems that students use to answer questions in class, have found that discussion among students (peer discussion) increases performance on these clicker questions (Mazur, 1997). Even without instructor input, student performance improves when students individually answer similar but new questions on a topic. This improvement occurs even when few students initially know the correct answer (Smith *et al.*, 2009) and is independent of overall achievement level of the individual students (Smith *et al.*, 2011). When instructors embrace such peer discussion and establish "sense-making" norms, for example, getting students to understand and articulate their reasoning, this behavior also impacts student attitudes about their learning experience. In such classes, students place a higher value on articulating their reasoning than students in classrooms in which the instructor focuses on getting correct answers (Turpen and Finkelstein, 2010). Thus, there is evidence to support both the idea that students are capable of constructing their own knowledge through peer discussion and that certain instructional practices positively affect student attitudes toward discussion.

Nonetheless, instructors who otherwise embrace using in-class activities sometimes hesitate to encourage small-group discussion due to various concerns, including losing classroom control, having discussions take too much time

DOI: 10.1187/cbe.13-05-0090
Address correspondence to: Jennifer K. Knight (knight@colorado.edu).

(requiring a reduction in content), and skepticism about whether students are capable of teaching themselves (Felder and Brent, 1996). Therefore, there remains a need to provide evidence that specific instructional methods can lead students into using in-class discussion time productively and that these discussions are more beneficial to students' development of skills and knowledge than lecturing.

Student discussions have been studied extensively in the K–12 classroom environment. Much of this work has focused on how students exchange ideas, disagree with one another, or support their ideas with reasons. In general, these interactions are referred to as "argumentation." Toulmin (1958) defined quality argumentation as a situation in which the speaker makes a claim, provides evidence or reasoning for his/her claim, and ultimately provides a warrant—a statement that links the initial claim to the supporting evidence. Toulmin's classification of the different elements of argumentation has been subsequently used and modified by others to help describe and characterize student dialogue (Driver *et al.*, 2000; Jiménez-Aleixandre *et al.*, 2000; Sampson and Clark, 2008). Some have focused on the correct use of scientific content (Sampson and Clark, 2008) or the frequency of rebuttals, in which a student challenges another student's initial offering of evidence (Osborne *et al.*, 2004), while others have explored how teachers' questions and prompts impact the nature of the discussion (Michaels *et al.*, 2002).

In terms of fostering student argumentation in the classroom, two conditions appear particularly important: an instructional task that challenges students to consider alternate ideas and a social context that invites dialogue (Osborne *et al.*, 2004). Many secondary school science classrooms do not provide opportunities to engage in argumentation (Lemke, 1990), and, when students are given the opportunity, they often do not readily employ reasoning in their arguments (Kuhn, 1991; Kuhn and Udell, 2003; Zohar and Nemet, 2002). However, when instructors use such behaviors as prompting students to use reasoning or modeling what such reasoning should look like, student argumentation skills of even young students improve, suggesting that argumentation is a skill that must be explicitly taught and practiced (Zohar and Nemet, 2002; Osborne *et al.*, 2004; McNeill *et al.*, 2006).

In contrast to the wealth of information on younger students, few studies have examined the content or nature of argumentation in college-level classrooms. When asked to construct a written argument to explain data, undergraduates in an introductory biology class were able to generate simple features of making a claim and using evidence but did not generally supply warrants for their reasoning or construct rebuttals unless explicitly directed to do so (Schen, 2012). They also struggled with providing alternative explanations for data, even when prompted. Similar patterns have been found in the analysis of oral argumentation. In a large introductory astronomy class, less than half of student clicker question discussions involved an exchange of claims and support or rebuttals of those claims with additional discussion (James and Willoughby, 2011). The majority of discussions involved considering ideas not presented in the clicker question or answers, showed a lack of understanding of basic principles needed to discuss the question, or went off task. In addition, some "discussions" mainly involved group members listening to the dominant person in the group rather than exchanging reasoning, especially when correct answers were rewarded with more points (also previously discussed in James, 2006). These studies suggest that college-level students are similar to younger students in their lack of argumentation skills and should benefit from explicit instruction in how to use discussion time productively.

If faculty members could be confident that certain techniques would lead their students to engage in meaningful discussions that improve learning, they might be more likely to implement such techniques. Thus, it is important both to further characterize the content of peer discussions in different types of undergraduate science classes and to identify ways that instructors can successfully influence the quality of student discussion. In this study, we have characterized the types of statements upper-level undergraduate biology students make when engaged in discussions of challenging conceptual clicker questions and have explored the relationship of these statements to the outcome of the discussions and whether explicit instructor cues affect the amount and kind of reasoning students use when given the opportunity to discuss.

METHODS

Characterization of the Course, Students, and Instructional Style

The students in this study were enrolled in an upper-division developmental biology course, one of several possible required capstone courses taken by majors in their junior or senior year. The course (taught by J.K.K.) was designed to challenge students to apply knowledge gained in four prerequisite core courses and to provide practice in critical thinking and problem solving. The course met for 75 min twice a week for 15 wk. The 107 students enrolled in the course (47% female, 97% majors) were seated at 12 round tables with 8–10 students per table. Students were allowed to self-select into groups, which were formed in the first week of class with encouragement from the instructor and maintained for the rest of the semester. The class was taught in an active, student-centered style: < 60% of class time was spent lecturing; students completed written in-class problem-solving exercises in at least 70% of class periods; and students were asked an average of five clicker questions per class period. Clicker questions were implemented following the peer instruction model (Mazur, 1997): students answered questions initially on their own; the instructor encouraged peer discussion (without showing the vote distribution) when less than 70% of students initially answered the question correctly; and students revoted after discussion. Following discussion and revote, volunteers were usually asked to share their ideas and problem-solving strategies with the rest of the class, and the instructor wrapped up discussion after showing the histogram distribution of the revote.

Instructor Cues

During the semester, the instructor varied how clicker questions were introduced prior to and following peer discussions, taking either an "answer-centered" or "reasoning-centered" approach (Table 1). The instructor alternated between these two kinds of cues on a weekly basis throughout the semester: for example, in one week, the instructor used

Table 1. Instructional behaviors in answer-centered and reasoning-centered class periods

	Answer-centered class	Reasoning-centered class
Instructor cue	"Discuss your answers with your table and revote. Then, I'll explain the correct answer."	"Discuss your answers with your table, and focus on the reasons for your answers. Then, I'll ask you to share your reasons."
Student reasoning requested	None	Volunteers were asked to share reasoning from their group's discussion.
Histogram of student answers	Shown immediately after discussion and revote	Shown after students volunteered reasoning
Instructor wrap-up	Instructor explained reasoning and correct answer.	Instructor highlighted student explanations that described correct answer.

the answer-centered approach in the two class periods, and, the next week, the instructor used the reasoning-centered approach. The instructor used slight variations in the cues, but always mentioned that she would explain the correct answer in the answer-centered cues and always mentioned that she wanted students to report on their reasoning in the reasoning-centered cues. No additional explicit instruction on how to exchange reasoning or how to construct an argument was provided. Other than the cue variations, the way materials were discussed, the overall active nature of the classroom (group exercises and frequent opportunities for discussion), and the nature of the clicker questions themselves did not differ among the class periods.

Subjects and Recordings

Four tables of students gave consent to be videotaped for the duration of the semester (University of Colorado IRB protocol #0603.08). The volunteer groups were representative of the class population in gender distribution (~50% male), attendance (one-way analysis of variance [ANOVA]: $p = 0.68$), and mean performance (one-way ANOVA: $p = 0.20$). Recordings were made during 17 class periods beginning in the third week of the semester. Audio was captured from several flat microphones placed between students at a table and routed through a mixer to a video camera. After the conclusion of the course, the discussions of clicker questions were transcribed verbatim from the videotapes. A maximum of five students could be reliably recorded at a time; thus, the transcribed discussions include only a subset of student voices from the 8–10 students typically seated at each table. However, because students usually shared their ideas table-wide before revoting, we analyzed student performance by table, not restricting these measures to only the students directly involved in the discussion.

Due to the nature of the recording equipment, discussions sometimes involved inaudible portions of conversation. However, recordings typically included a logical thread of conversation among two to five students. Discussions that were mostly inaudible were not included in this study.

Analysis: Fine-Grained Discussion Coding

We identified each student statement as a turn of talk—if a student spoke several sentences in a row, this was still defined as one turn of talk. We then gave each turn within a transcript one or more codes representing its role in the discussion. We tracked several elements of discussion: claim making (stating their choice of an answer), explaining reasoning for an answer, asking a question, or providing background information (Table 2). Statements that were off topic or simple statements of agreement or disagreement ("Okay," or "That doesn't make sense") were given a code of NA (not applicable). Turns of talk were given more than one code if they included more than one element (e.g., a statement might have been coded as both a question and a background statement), or if they included more than one instance of the same element (e.g., multiple pieces of evidence used to support a reason). The coding scheme was determined and refined by four coders; after an interrater agreement of greater than 0.7 on 10% of the transcripts was achieved, the remaining transcripts were independently coded. The total turns of talk were calculated for each discussion, as was the fraction of the discussion devoted to each of the codes described above.

Analysis: Whole-Discussion Measures

In addition to fine-grained coding, each transcript was categorized on a whole-discussion level for several different features. After considering other whole-discussion coding

Table 2. Definition of fine-grained codes[a]

Code	Definition
Claim	A statement of preference for an answer (such as "I think it's 'A'")
Reasoning	A unique explanation for choosing or eliminating an answer, including warrants and rephrasings of previously given reasons
Question	Any question (asking peer to explain idea, asking about definitions, wording, or background information)
Background	Providing information about a question to clarify what the question was asking or what a figure showed
NA	Any speech not described by the above codes, including simple statements of agreement or disagreement, reasoning statements repeated verbatim, joking, and off-topic talk

[a] Each turn of talk was given one or more codes as defined.

Table 3. Exchange of Quality Reasoning levels[a]

Level	Definition
0	No students made reasoning statements.
1	Only one student used reasoning, which could include a warrant (no exchange).
2	Two or more students exchanged reasoning, but neither or only one included warrants.
3	Two or more students exchanged reasoning, including warrants.

[a] Each transcript was assigned a level based on the characteristics described.

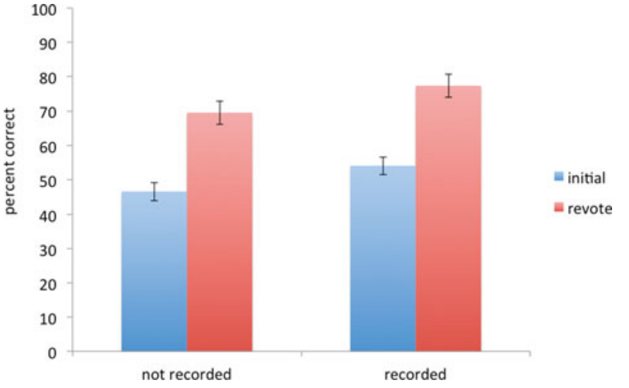


Figure 1. Average percent correct on transcribed clicker questions for unrecorded (average $n = 60$) and recorded (average $n = 35$) students. The initial and revote values are not significantly different between the two groups ($p > 0.05$; two-way repeated measures ANOVA). Revote values are significantly higher than initial for both groups ($p < 0.001$; two-way repeated measures ANOVA, no interaction $p = 0.88$). Error bars indicate SEM.

schemes (e.g., Osborne *et al.*, 2004), we developed a scoring system to measure the quality of student reasoning (Exchange of Quality Reasoning; Table 3) based on exchanges of warrants (Toulmin, 1958). The highest-scoring discussions, level 3, feature multiple students linking evidence to a claim with logical reasoning, while the lowest-scoring discussions involve no exchange of reasoning although they usually involve exchanges of claims. Each discussion was scored independently by J.K.K. and S.B.W. (interrater agreement of 0.86); any differences were resolved by consensus. We also tracked discussions containing conflicting lines of reasoning, in which one student's reason was challenged by at least one other student's reason.

Several additional whole-discussion measures were taken: the amount of time spent in discussion, the percent correct on the initial and revote for all students at each recorded table, and the normalized change in voting for each recorded table of students discussing a clicker question. The normalized change formula uses the original normalized gain formula (Hake, 1998) of: 100(revote − initial)/(100 − initial) for positive changes, and the formula: 100(revote − initial)/initial for negative changes (Marx and Cummings, 2007). If the percent correct was 100% for both initial and revote, the discussion was not included in the average calculation.

RESULTS

Characterizing Discussions

General Characterization of Clicker Questions and Clicker Discussions. Over the semester, the instructor posed a total of 124 clicker questions. When the initial vote was < 70% correct, the instructor asked students to discuss their answers and revote: 81 questions fell into this category. Each discussion lasted an average of 2 min (± 0.93 SD) and involved from two to five students. During each class in which recordings were made, one to three clicker questions were selected at the end of each class period by the instructor for transcription, based on the clarity of the questions (one correct answer) and their likelihood of generating discussion. After listening to these recordings, we had to exclude some due to poor audio quality. Ultimately, discussions of 34 questions from the 17 class periods in which recordings were made were transcribed and further analyzed. Multiple groups were recorded discussing each question, resulting in a total of 83 transcripts. In most cases, the student votes at a particular table were similar to the class-wide average of < 70% correct. In 17 recorded discussions, the initial vote for students at one or more tables was higher than 70%; these discussions were still transcribed and analyzed.

The 34 clicker questions selected for analysis were rated as either requiring Bloom's lower-order cognitive skills (LOCS) or higher-order cognitive skills (HOCS; Crowe *et al.*, 2008) by two independent raters and the instructor; a final rating for each question was reached by consensus. Twenty-nine of the questions were rated as requiring HOCS, and five as requiring LOCS.

The 34 transcribed clicker questions were on average more difficult than the complete set of clicker questions for the course (48% correct vs. 61% correct, respectively), because discussions and revotes were only performed when the overall vote was < 70% correct. The students who volunteered to have their discussions recorded did not perform significantly differently from the rest of the students in the class on clicker questions (Figure 1) or on other measures, such as exam performance, suggesting that the volunteers were representative of the class. After peer discussion, the average percent correct on the revote was significantly higher than the average initial vote for both recorded and unrecorded students, as has been reported for other studies on peer discussion (e.g., Smith *et al.*, 2009, 2011).

Transcript Coding

Fine-Grained Coding. For each transcript, each student turn of talk was characterized using the codes shown and explained in Table 2. The total turns of talk were summed for each discussion and were used to characterize the percent of turns of talk devoted to each feature. On average, students spent a larger percent of each discussion explaining their reasons to each other than making claims, asking questions, providing background, or making NA remarks (Table 4), although there was considerable variability among discussions.

To determine whether the features of each discussion were affected by the initial percent of students who answered each clicker question correctly, we looked for correlations between the initial percent correct and the percent of turns of talk spent

Table 4. Prevalence of each code within clicker discussions[a]

Code	Average percentage of turns of talk for all discussions ($n = 83$)
Claim	30.8 (1.9)
Reasoning	39.2 (2.4)
Question	17.7 (1.0)
Background	5.7 (1.0)
NA	28.3 (1.7)

[a] Each transcript was treated as a single discussion. Discussions had on average 20.5 (1.2) turns of talk. Values do not total 100%, because each turn of talk could be given more than one code. SEM shown in parentheses.

on each of the fine-grained coding features (Tables 2 and 4). No significant correlations exist (Pearson's r values: all <0.15; $p > 0.05$). For example, in transcripts for which students began with a low initial percent correct, some discussions devoted a high percentage of the discussion to reasoning, and others a low percentage. In the 17 discussions for which the initial vote for that group of students was above 70% correct, the students still spent on average 42% of the discussion explaining their reasoning, similar to the overall average for all discussions (39%).

To determine whether the percent of turns of talk spent on these features subsequently impacted the percent of students who answered correctly on the revote, we also looked for correlations between these two measures. As was true for the initial vote, no correlations exist between individual features and the percent correct on revote (Pearson's r values: all <0.18; $p > 0.05$). Discussions in which a high proportion of the discussion was spent exchanging reasoning or a high proportion was spent exchanging claims were no more likely to produce a high correct revote than were discussions in which a low proportion of time was spent on reasoning or claims. Thus, the fraction of discussion spent on different features of talk is not necessarily predictive of the vote outcome.

Finally, there was no correlation between time of the semester and each of the measures reported above; although individual discussions vary in student use of reasoning and in student performance, there are no positive or negative trends from beginning to end of the semester (Pearson's r values: all <0.02; $p > 0.05$).

Whole-Discussion Coding of Reasoning. Reasoning statements for all discussions were coded as described above. Some reasoning statements were consensus building, in which a student agreed with a previous student, and added to the line of reasoning. Some reasoning statements were conflict oriented, in which a student disagreed with a line of reasoning previously made and gave a conflicting reason to support his or her claim. Willingness to engage in disagreements with one another was quite common: 66% of the discussions included conflicting reasoning statements.

To characterize the reasons that students provided, we scored each of the discussions using the Exchange of Quality Reasoning measure (Table 3). This measure emphasizes the value both of exchanging reasoning and of providing reasoning that logically connects evidence with a claim (warrants).

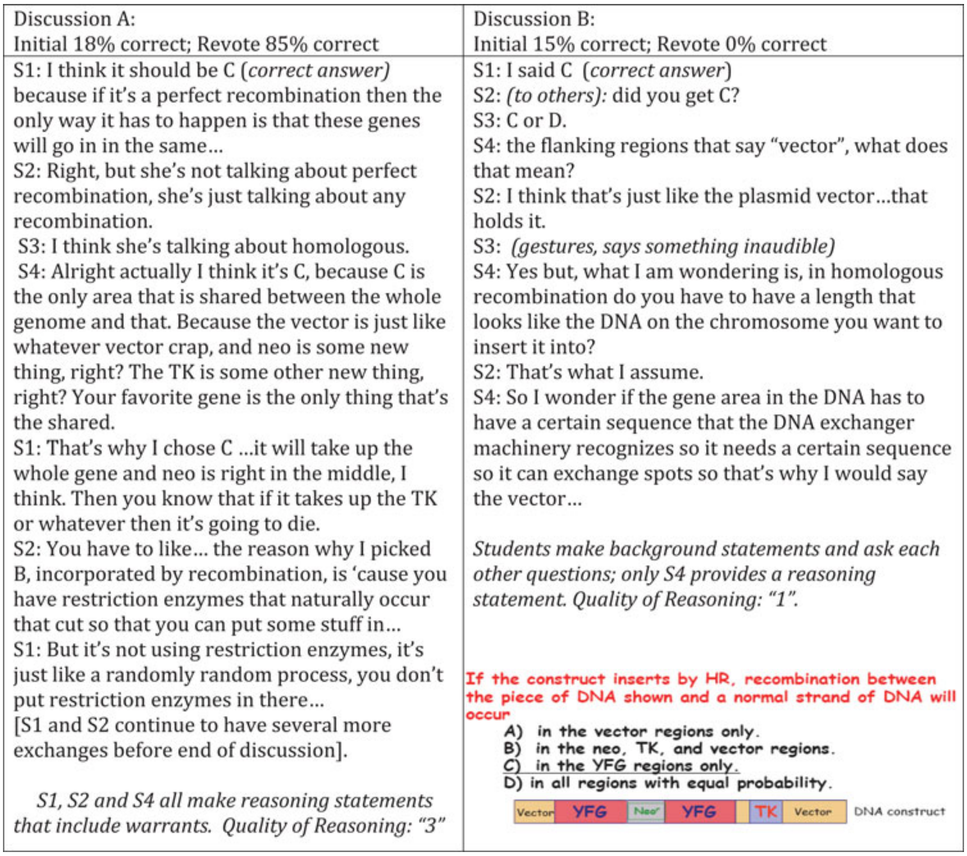
An example of two discussions that illustrate the difference between a level 1 and a level 3 score for Exchange of Quality Reasoning is shown in Figure 2. The discussions are about the same clicker question, and both began with a low percent correct initial vote (15–18%). In discussion A, students used many reasoning statements to support their claims, and their reasoning statements included warrants connecting their reasons with evidence. In discussion B, only one student gave a reasoning statement, while the rest of the students asked questions and made background comments. Discussion A was scored a 3 and resulted in 85% of the students at the table answering the question correctly in the revote; discussion B was scored a 1, and 0% of the students at the table answered the question correctly in the revote. Most of the 83 discussions were categorized into the two highest-quality reasoning levels, both of which involve exchanges of reasoning (level 3: 54%; level 2: 24%). A smaller number involved reasoning provided by one student only (level 1: 18%), and only three discussions exhibited no reasoning at all (level 0: 4%). Because so few discussions scored "0" level, we combined discussions from levels 0 and 1 for additional analysis.

To examine whether discussions scored at different levels of reasoning resulted in different learning outcomes, we compared the mean percent correct on revotes and the normalized change in correct voting that followed discussions of each level (Figure 3). Discussions that involved an exchange of warrants (level 3) had the highest normalized change and the highest percent correct on revotes. However, correct revoting was not significantly different between level 3 and the other levels of reasoning. As it is not valid to apply statistical tests to normalized change values (Marx and Cummings, 2007), only the SE bars for each value are displayed in Figure 3.

To assess whether any fine-grain coded features reported above are associated with discussions that score highly for reasoning quality, we tested the relationship between the frequency of a discussion feature and its quality-of-reasoning level. Discussions that involved an exchange of reasoning (levels 2 and 3) devoted a significantly higher percent of the discussion to reasoning and also had significantly more turns of talk, compared with level 0/1 discussions (Table 5). In other words, when students exchanged ideas, they engaged in longer discussions and provided a significantly greater number of statements of reasoning than in scenarios in which only one person was offering reasons for his or her answer.

Finally, we investigated whether there was a correlation between Bloom's level of a question and the quality of reasoning that students used when discussing the question. The discussions in the 0/1 quality of reasoning level were distributed over 15 of the 34 clicker questions. Three of the 15 questions were judged as requiring LOC- rather than HOC-level skills and resulted in five of the 18 level 0/1 discussions, suggesting that these particular questions were less rich for generating discussion than other questions in the data set. However, the two other LOC clicker questions generated level 2/3 discussions in all recorded groups.

Differences between Groups of Students. We recorded from four different tables of students over the course of the semester. One table chose to exit the study after a few weeks; we therefore recruited another table to take its place. Thus, the number of transcripts obtained was not equal among the tables of students, and it was necessary to investigate whether table identity affected the results. We found that tables did not differ significantly in the distribution of quality-of-reasoning



Discussion A: Initial 18% correct; Revote 85% correct	Discussion B: Initial 15% correct; Revote 0% correct
S1: I think it should be C (*correct answer)* because if it's a perfect recombination then the only way it has to happen is that these genes will go in in the same… S2: Right, but she's not talking about perfect recombination, she's just talking about any recombination. S3: I think she's talking about homologous. S4: Alright actually I think it's C, because C is the only area that is shared between the whole genome and that. Because the vector is just like whatever vector crap, and neo is some new thing, right? The TK is some other new thing, right? Your favorite gene is the only thing that's the shared. S1: That's why I chose C …it will take up the whole gene and neo is right in the middle, I think. Then you know that if it takes up the TK or whatever then it's going to die. S2: You have to like… the reason why I picked B, incorporated by recombination, is 'cause you have restriction enzymes that naturally occur that cut so that you can put some stuff in… S1: But it's not using restriction enzymes, it's just like a randomly random process, you don't put restriction enzymes in there… [S1 and S2 continue to have several more exchanges before end of discussion]. *S1, S2 and S4 all make reasoning statements that include warrants. Quality of Reasoning: "3"*	S1: I said C (*correct answer*) S2: *(to others):* did you get C? S3: C or D. S4: the flanking regions that say "vector", what does that mean? S2: I think that's just like the plasmid vector…that holds it. S3: *(gestures, says something inaudible)* S4: Yes but, what I am wondering is, in homologous recombination do you have to have a length that looks like the DNA on the chromosome you want to insert it into? S2: That's what I assume. S4: So I wonder if the gene area in the DNA has to have a certain sequence that the DNA exchanger machinery recognizes so it needs a certain sequence so it can exchange spots so that's why I would say the vector… *Students make background statements and ask each other questions; only S4 provides a reasoning statement. Quality of Reasoning: "1".* If the construct inserts by HR, recombination between the piece of DNA shown and a normal strand of DNA will occur A) in the vector regions only. B) in the neo, TK, and vector regions. C) in the YFG regions only. D) in all regions with equal probability.

Figure 2. Transcripts of two different discussions of the same clicker question (shown at bottom). The initial and revote percents correct for each group are shown, as well as the quality of reasoning level assigned to the transcript. Individual students in each group are indicated as S1, S2, etc.

levels in their transcripts (Kruskal Wallis: $p = 0.07$); the fraction of discussions that included conflicting reasoning (chi-squared: $p = 0.55$); or the measures of initial percent correct, revote percent correct, or normalized learning change (one-way ANOVA: $p > 0.05$). There do exist differences in some of the fine-grained coding measures: one table devoted more of their discussions to claims, while a different table spent more of their discussions on reasoning than did the other tables (one-way ANOVA, Tukey post hoc: $p < 0.05$). Because the discussions of each table differed significantly on very few measures, and within-table variability exceeded between-table variability, analyses of discussions reported above were done with all tables aggregated.

Further qualitative analysis of the discussions of different tables revealed some interesting features. One table was especially collaborative: students added to one another's ideas, collectively building on an often correct but incomplete idea to ultimately create a well-supported reason for their answer. When they did initially disagree on the correct answer, they provided evidence to support their conflicting reasons, exchanging ideas until they arrived at an answer that made sense to the group as a whole. Although not statistically significant, this group had the highest percent correct both on their initial votes (60.5%) and on their revotes (87.5%), suggesting that despite an initially high understanding, they readily engaged in further discussion, helping others at the table to achieve understanding. Another table had a different style of interacting: they made more jokes and were more (good-naturedly) critical of one another. This table made the most claim statements and often disagreed on their initial vote. Although not significantly different from other tables, their discussions frequently contained conflicting pieces of evidence to support their claims. They had a lower initial percent correct than the previous table (54.5%) but still a reasonably high average percent correct on revote (74.5%). Both these tables performed slightly but not significantly better with respect to revote percent correct than the other two tables, which spent more time expressing frustration and uncertainty or listening to one person who appeared to have an idea of the correct answer.

Impact of Instructional Cues. To measure the impact of different instructional cues on student discussion, we separated discussions into two categories: those following reasoning-centered cues (51 discussions), and those following answer-centered cues (32 discussions; Table 1). Discussions in the two categories did not differ in the average time spent discussing clicker questions, the average percent correct on initial vote, the average percent correct on revote, or the fraction of the discussion spent on reasoning (Table 6), paralleling our finding that many discussion features are not directly correlated with measures of performance. However, when the instructor

Table 5. Characteristics of discussions scored by Exchange of Quality Reasoning

Exchange of Quality Reasoning level	Number of discussions	Average turns of talk per discussion (SEM)	Average percent of discussion devoted to reasoning (SEM)	Average percent correct on revote
3	45	24.7 (1.7)	44 (2.7)	79.8
2	20	20 (1.6)	43 (6.0)	69.1
0/1	18	10.7 (1.0)[a]	22 (3.2)[a]	70.5

[a] Level 0/1 discussions significantly lower than levels 2 and 3; $p < 0.05$ (one-way ANOVA). Level-2 and level-3 discussions were not significantly different from each other on any of these measures.

used reasoning cues, students engaged in significantly more high-quality discussions that included exchanges of warrants (level 3) than when the instructor cued students to focus on the answer. In turn, the fraction of the discussion spent on claims was significantly lower in reasoning-cued discussions. Reasoning-cued discussions were also more likely to exhibit conflicting lines of reasoning among students (73%) than were answer-cued discussions (56%), although this difference is not statistically significant (Table 6).

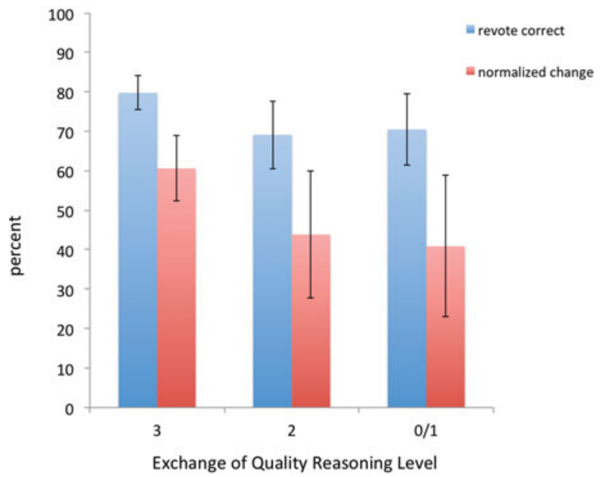


Figure 3. Outcome measures for tables of students, by Exchange of Quality Reasoning level (levels 0 and 1 combined, $n = 18$; level 2, $n = 20$; level 3, $n = 45$). The mean percent correct on revotes for each set of scored transcripts is shown in blue (no significant differences between levels; one-way ANOVA, $p > 0.05$). The mean percent normalized change for each set of scored transcripts is shown in red. Bars indicate SEM.

DISCUSSION

In this study we characterized the features of high- and low-quality peer discussions of in-class clicker questions among upper-division undergraduate biology majors. We analyzed how the features of these discussions related to performance, and we discovered that certain features of discussion differ in response to instructor cues.

Upper-Division Students Generally Engage in Productive Discussion

We find that students in general vote more correctly following peer discussion, supporting previous work (Smith *et al.*, 2009, 2011) and indicating that their engagement in peer discussion improved their understanding (Figure 1). In contrast to introductory astronomy students (James and Willoughby, 2011), recorded volunteers in this upper-division course engaged in the type of discussion the instructor intended for almost all of the transcripts analyzed: that is, they exchanged reasoning related to the clicker question asked. In only three cases did students fail to discuss their ideas after exchanging information about their votes.

Smith *et al.* (2009) suggested that improvement of student performance on clicker questions likely results from a

Table 6. Comparison of answer-cued and reasoning-cued discussions[a]

	Answer cued ($n = 32$)	Reasoning cued ($n = 51$)
Time (minutes)	2.5 (0.5)	2.7 (0.8)
Turns of talk	18.7 (1.4)	21.7 (1.7)
Percent correct initial vote	56.9 (3.6)	48.8 (3.6)
Percent correct revote	80 (5.3)	72 (4.9)
Percent devoted to claims	36.2 (3.4)*	27.4 (2.1)
Percent devoted to reasoning	33.4 (3.4)	42.8 (3.2)
Exchange of Quality Reasoning level	2.0 (0.1)	2.5 (0.1)**
Percent of discussions that involved conflicting lines of reasoning	56	73

[a] Average time, turns of talk, and percent correct initial and revote were not significantly different between the two conditions (independent samples t test: $p > 0.05$ in all cases). The percent of discussions that involved conflicting lines of reasoning was also not significant (Mann-Whitney U-test: $p = 0.129$). All values shown are averages. SEM is shown in parentheses.

* $p < 0.05$, statistically significant difference between answer-cued and reasoning-cued discussion (independent samples t test).

** $p < 0.01$, statistically significant difference between answer-cued and reasoning-cued discussion (Mann-Whitney U-test).

cooperative group construction of knowledge (coconstruction), rather than simply from one student telling other students the correct answer (transmission). We found evidence of coconstruction in more than three-quarters of these discussions, which exhibited either an exchange of reasoning (level 2) or an exchange of warrants (level 3). It is possible that this comparatively high-level reasoning behavior may be explained by the students' upper-division status. These students had used clickers in most of their core biology courses and likely had opportunities to develop sophistication in reasoning, particularly in so-called critical-thinking courses, which involve reading and presenting on original literature. Presumably, most upper-division biology majors would also have developed at least some of these skills through their course work. Finally, the students in this study may have further developed reasoning skills specifically during this course as they became more familiar with answering and discussing challenging clicker questions. If this development occurred, it likely had the largest impact prior to the onset of recording in week 3 of class, because we did not see evidence of an improvement in quality of reasoning over the semester. However, such a change as a result of in-course practices may occur in less advanced students, such as those enrolled in introductory biology, who are the target audience of our future studies.

The Bloom's Level of a Question Does Not Necessarily Influence Student Discussion

Answering most clicker questions in this study required Bloom's HOC skills. We found that the five questions rated as requiring LOC skills still had the potential to generate student discussion that involved exchanges of reasoning. This finding is consistent with James and Willoughby's work (2011), in which the authors reported that introductory astronomy students discuss "recall" (Bloom's level 1) questions extensively, despite instructors' perceptions that these questions are simple or basic. Thus, the cognitive level of a question does not necessarily correlate with its perceived easiness or difficulty as judged by instructors (Lemons and Lemons, 2013) and does not determine the quality of the discussion among students.

Initial Votes on Clicker Questions Do Not Determine Discussion Features

In this course, when the class-wide vote was above 70% correct, the instructor did not have students discuss and revote on the question. However, recorded discussions among groups of students varied—sometimes the initial vote at a table was above 70% correct, even though the class-wide vote was below 70%. Thus, we had the opportunity to investigate how students discussed questions when many of them were already in agreement about the correct answer. We found that, when students already had the correct answer, they still discussed their ideas and were just as likely to exchange claims, questions, and warrants as when they did not already have the correct answer. This may suggest that, contrary to previous assumptions, upper-division students find value in pursuing discussion even when many of them already have voted for the correct answer. Some possible reasons for this behavior include: 1) students may initially vote for the correct answer without feeling confident in their explanation for that answer, or 2) students may vote for the correct answer but may have the wrong reasoning to support this answer. Additionally, because this class involved up to 10 students sitting at the same table and working together as a group, it is likely that even a single difference of opinion, if voiced, could spur discussion.

Why Do Students Sometimes Fail to Perform Better after Discussion?

In 25% of the discussions analyzed, individual tables of students did not improve after discussion; in some cases, the percent correct on the revote remained the same, while in other cases, more students selected the incorrect answer than during the individual vote. Even when more than 70% of the class achieved the correct answer after discussion, there were sometimes whole tables of students who did not revote correctly. Reviewing discussions of this nature revealed that this scenario often occurs when one or more students begin with an incorrect idea and are able to use evidence to convincingly support this idea, even though the evidence is not factually correct. In other instances, one student with the correct idea may not supply convincing reasoning or may supply no reasoning at all, in which case, the other students stick with their initial incorrect answer. Finally, in some cases, none of the students have the correct idea, no one is motivated to share his or her reasoning, or no support for the correct idea is offered within a group, leading to no change in the students' ideas (as evidenced by discussion B in Figure 2). In summary, students who vote incorrectly in this situation are not necessarily isolated individuals spread out in the classroom who did not participate in the discussion. These findings support the recommendation that whole-class discussion of both the most commonly chosen incorrect answer and the correct answer should be helpful for students, even when most students have answered the question correctly (Caldwell, 2007). It also further supports the practice of not showing the histogram of student answers until after students have an opportunity to share their reasoning with the class, so students are not biased in their discussion by the majority vote (Perez *et al.*, 2010).

Instructor Cues Influence Quality of Student Discussion

This course was taught in a student-centered, active style throughout the semester. Class periods were very similar in terms of expectations for student participation and the engagement of students in other in-class activities besides clicker questions. The only appreciable difference in the answer-cued and reasoning-cued class periods was how the clicker question discussions were cued, and how the instructor followed through with class-wide discussion (Table 1). Interestingly, focus group interviews with volunteers who had participated in the recordings revealed that students were not explicitly aware of the different cueing. Thus, even though the cues used involved subtle changes in the patterns of instructor–student interaction, student discussion behavior still shifted significantly in response to differences in these cues.

We propose several possible explanations for the increased quality of reasoning in the reasoning-cued discussions. One possibility is that the students are responding solely to the cue: they hear the suggestion to use reasoning, and this reminder is enough to stimulate such discussion. Another possibility is that students are reacting to the instructor's cue that they will be held accountable for their discussion: accountability has been shown to impact student reports of their attentiveness to the task at hand (Nichol and Boyle, 2003). In reasoning-cued discussions, students were told they would be asked to explain their tables' reasons to the rest of the class. This may have motivated students to focus on being able to explain their ideas, thus encouraging them to provide evidence for their claims. Supporting this as a potential mechanism, we noticed several instances in the reasoning-cued discussion of students expressing concern that they be able to explain their ideas: "If she calls on me, I'll die because I have no idea," and "But if she asks how do we know we're right, how do we know it's the right answer?" Finally, another possibility is that the students are *negatively* affected by answer-cued instruction. By placing emphasis on achieving a correct answer and by leading students to expect an instructor explanation, students may be prevented from engaging in their "normal" level of discussion. Thus, although accountability may in fact be a strong motivator, further work is necessary to exclude the possibility that answer cueing is demotivating.

Value of Student Discussion

We find it notable that exchanging reasoning does not guarantee that students will arrive at a correct answer: the fraction of each discussion spent explaining reasoning did not correlate with the percent of students at a given table who ultimately answered the question correctly (Table 5). As most discussions contained reasons supporting both correct and incorrect answers, it is not surprising that students sometimes led each other in an incorrect direction. On the other hand, discussions that included exchanges of warrants of those reasons, in which evidence was used to explain a claim or justify an idea, resulted in a higher percent of students answering correctly (Table 5 and Figure 3). Although these were not significant differences, the findings suggest a tendency for the highest-quality discussions to result in more correct answers.

Nevertheless, these trends do not address what we think is a critical benefit of student discussion: the act of discussion itself is an important component of learning, regardless of whether students are immediately able to select the correct answer for a clicker question. Socially mediated communication has been shown to be crucial for an individual's exposure to and practice with new ideas, and is a frequent prerequisite to individual internalization of concepts (Vygotsky, 1978; Lave and Wenger, 1991). Science content learning is mediated through language and communication, and, in the education process, students have the opportunity to acquire the language of science through discourse (Osborne, 2010). In addition, dialectical argumentation, in which more than one side of an argument is explored, has also been shown to be central in the learning process (Asterhan and Schwarz, 2009). Thus, even when students argue for incorrect answers, they are engaging in this process of learning. In addition, anecdotally, students often refer back to previous discussions of clicker questions. We have observed this behavior both when students are considering new clicker questions and in other problem-solving settings, such as help sessions and homework-solving sessions. This suggests that students are remembering and using previous peer discussions to help them reason through new scenarios.

In summary, we have shown that upper-level biology students in a student-centered course readily discuss their answers to clicker questions by exchanging reasons and providing evidence for their ideas. We have also demonstrated that the initial vote or clicker question type does not determine the amount or quality of the reasoning, suggesting that students can benefit from discussion no matter what the conditions. In addition, we show that students follow numerous paths in their discussion: paths that consider multiple answers and result in a correct answer, as well as paths that lead to an incorrect answer due to exchange of incorrect reasoning, one convincing person with an incorrect idea, or an absence of discussion. Most importantly, the evidence presented in this paper supports a critical role for the instructor in stimulating high-quality discussions of clicker questions. Students changed their discussion behavior in response to instructor cues, using more quality reasoning when the instructor emphasized using and sharing reasoning. Whether these discussions not only help students with the social process of problem solving and understanding material in class, but also impact retention and understanding of concepts long term, deserves further study.

REFERENCES

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

Caldwell JE (2007). Clickers in the large classroom: current research and best-practice tips. CBE Life Sci Educ *6*, 9–20.

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.

Driver R, Newton P, Osborne J (2000). Establishing the norms of scientific argumentation in classrooms. Sci Educ *84*, 287–312.

Felder RM, Brent R (1996). Navigating the bumpy road to student centered instruction. Coll Teach *44*, 43–47.

Freeman S, O'Connor E, Parks JW, Cunningham M, Hurley D, Haak D, Dirks C, Wenderoth MP (2007). Prescribed active learning increases performance in introductory biology. CBE Life Sci Educ *6*, 132–139.

Hake RR (1998). Interactive-engagement vs. traditional methods: a six-thousand-student survey of mechanics. Am J Phys *66*, 64–74.

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.

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

Kitchen E, Bell JD, Reeve S, Sudweeks RR, Bradshaw W (2003). Teaching cell biology in the large-enrollment classroom: methods to promote analytical thinking and assessment of their effectiveness. CBE Life Sci Educ 2, 180–194.

Kuhn D (1991). The Skills of Argument, Cambridge, UK: Cambridge University Press.

Kuhn D, Udell W (2003). The development of argument skills. Child Dev *74*, 1245–1260.

Lave J, Wenger E (1991). Situated Learning: Legitimate Peripheral Participation, Cambridge, UK: Cambridge University Press.

Lemke JL (1990). Talking Science Language, Learning, and Values, Norwood, NJ: Ablex.

Lemons PP, Lemons JD (2013). Questions for assessing higher-order cognitive skills: it's not just Bloom's. CBE Life Sci Educ *12*, 47–58.

Marx J, Cummings K (2007). Normalized change. Am J Phys *75*, 87–91.

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

McNeill KL, Lizotte DJ, Krajcik J, Marx RW (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. J Learn Sci *15*, 153–191.

Michaels S, O'Connor C, Hall M, Resnick L (2002). Accountable Talk: Classroom Conversation That Works, Pittsburgh, PA: University of Pittsburgh.

Nichol DJ, Boyle JT (2003). Peer instruction versus class-wide discussion in large classes: a comparison of two interaction methods in the wired classroom. Stud High Educ *28*, 457–473.

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.

Perez KE, Strauss EA, Downey N, Galbraith A, Jeanne R, Cooper S (2010). Does displaying the class results affect student discussion during peer instruction? CBE Life Sci Educ *9*, 133–140.

Sampson V, Clark DB (2008). Assessment of the ways students generate arguments in science education: current perspectives and recommendations for future directions. Sci Educ *92*, 447–472.

Schen M (2012). Assessment of argumentation skills through individual written instruments and lab reports in introductory biology. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, held 25–28 March 2012 in Indianapolis, IN.

Smith MK, Wood WB, Adams WK, Wieman C, Knight JK, Guild N, 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.

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

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.

Udovic D, Morris D, Dickman A, Postlethwait J, Wetherwax P (2002). Workshop biology: demonstrating the effectiveness of active learning in an introductory biology course. BioScience *52*, 272–281.

Vygotsky LS (1978). Mind in Society, Cambridge, MA: Harvard University Press.

Walker JD, Cotner SH, Baepler PM, Decker MD (2008). A delicate balance: integrating active learning into a large lecture course. CBE Life Sci Educ *7*, 361–367.

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