Analysis of student engagement in an online annotation system in the context of a flipped introductory physics class

Kelly Miller,[1] Sacha Zyto,[2] David Karger,[2] Junehee Yoo,[3] and Eric Mazur[1]
[1]Department of Physics and Division of Engineering and Applied Sciences, Harvard University, 9 Oxford Street, Cambridge, Massachusetts 02138, USA
[2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Building 32 Vassar Street, Cambridge, Massachusetts 02139, USA
[3]Department of Physics Education, Seoul National University, Seoul 151-742, South Korea
(Received 10 October 2015; published 14 December 2016)

We discuss student participation in an online social annotation forum over two semesters of a flipped, introductory physics course at Harvard University. We find that students who engage in high-level discussion online, especially by providing answers to their peers' questions, make more gains in conceptual understanding than students who do not. This is true regardless of students' physics background. We find that we can steer online interaction towards more productive and engaging discussion by seeding the discussion and managing the size of the sections. Seeded sections produce higher quality annotations and a greater proportion of generative threads than unseeded sections. Larger sections produce longer threads; however, beyond a certain section size, the quality of the discussion decreases.

DOI: 10.1103/PhysRevPhysEducRes.12.020143

I. INTRODUCTION

It is generally accepted that students understand material better after discussing it [1,2]. Discussion forums have been used successfully as tools to facilitate interactions and exchanges of knowledge between learners and between learners and instructors [3–6]. The social constructive theory of learning with technology [7] emphasizes that successful learning requires continuous conversation between learners as well as between instructors and learners [8]. The asynchronous nature of online discussion forums allows for discussion between learners and between learners and instructors at any time of day or night. This is a major advantage over other forms of communication [8].

Other advantages of online discussion forums include greater student participation, enhanced academic performance, and increased opportunity for metacognition [9,10]. Students have been found to participate significantly more in online discussions than in the traditional classroom [9]. In a study on the effect of computer-mediated discussions on academic performance, it was found [11] that students who engaged actively in the online discussions performed better in the course than other students [11]. Online discussion forums have also been shown to promote an increased exchange of ideas [12], an improved ability to make connections between concepts, and to apply the course material to diverse contexts [9]. Online discussion forums allow students more time to think and process responses to questions than during a regular face-to-face discussion [13,14]. Work on computer-mediated collaborative learning environments has shown that these tools can serve to scaffold online peer interactions through prompting strategies which have been shown to help students develop scientific thinking skills for asking and answering questions [15,16].

Despite the proclaimed benefits of online discussion forums, studies have shown that the value of online discussion forums is highly dependent on both the manner in which they are integrated into the learning environment and how they are tied to the assessment of the class. Online discussions have been shown to support the classroom discussion only when they encourage students to share different interpretations and perspectives of the course material and develop understanding of the material through debate [17]. Other studies have found online discussions that do not promote higher levels of thinking are ineffectual in providing increased learning [18].

Other studies have emphasized the importance of assessment for successfully implementing online discussion tools. Participation in online discussion forums is more active when it is linked to assessment [5]. Assigning a grade for participation in the discussion is necessary to ensure that students take part [19]. Despite the increasing use of forums in both online and residential courses, there is little existing work on how participation in online discussion facilitates learning and which measures can be taken to increase student participation.

In this paper, we study student participation in an online annotation system that allows students to discuss the reading online. We explore ways to increase the quality of the discussion to produce high-quality learning

interactions. In the context of a flipped, introductory physics class, we investigate three questions: What is the relationship between students' participation in the online discussion and their performance in the course? What effect does seeding the discussion with comments from previous iterations of the course have on student participation in the discussion? What effect does varying the number of students in a section have on student participation in the discussion?

II. STUDY

A. Course description

We conducted this study in the School of Engineering and Applied Science at Harvard University. The physics course, Applied Physics 50 (AP50), is an introductory, calculus-based physics course intended for engineering and premedical students. AP50 is split into two courses, AP50A (mechanics), which is taught in the fall semester and AP50B (electricity and magnetism), taught in the spring. We collected reading assignment data over two semesters of AP50 (number of students, $N = 91$ and $N = 70$), during the fall of 2013 and the spring of 2014. More than half of the students ($N = 57$) from AP50A took the AP50B class in the second semester. All the students who did not take AP50A in the fall took an alternative, equivalent mechanics class at Harvard. We determined that the students who took AP50B without having taken AP50A were not statistically different on any performance metric (precourse conceptual survey, average exam score, in class ConcepTest performance) from the students who took AP50A. We also determined that students who did not take AP50A did not differ in the quality of their annotations. Because the same instructor taught both semesters, the same pedagogy was used and the two student populations were determined to be the same on all metrics used in the analysis, we pooled the data from both semesters for analysis.

B. Pedagogy

Applied Physics 50 is taught using a flipped classroom approach. The class meets twice weekly for 3 h. There are no additional sections or labs; all of the course components are contained within this time. The pedagogy draws on features from both Project Based Learning [20] and Team Based Learning [21]. Project Based Learning is a teaching strategy in which students work for an extended period of time on an inquiry-driven project, often inspired by a real-world problem. By researching and problem solving, students gain knowledge and skills in specific content areas. All of the learning goals for AP50 are addressed through three projects per semester that students work on in class, as part of a team. Team Based Learning is a teaching strategy that has students organized into small, permanent groups. Students work within these groups for all aspects of the course, including assessments, which have both an individual and team component.

There are six different types of in-class activities, each of which is described below. In-class activities are designed to help students master the relevant physics and get started on the projects, which serve as the focal point for the course. The class activities are interrelated and are scheduled in an order that provides scaffolding as students learn the concepts. The schedule for the spring course (see Supplemental Material [22]) shows how the 6 different activities are structured into each 3 h class. The distribution and organization of activities in the fall semester was the same as in the spring. Each class contains anywhere from 1 to 3 activities with the more structured, instructor-led activities (Peer Instruction, for example) at the beginning of the class. As there are no lectures, students are expected to read the textbook online and develop a certain level of comfort with the material before coming to class. All students are exposed to the same activities and projects in the same sequence.

C. In-class activities

Peer Instruction: over the course of each semester, the instructor conducts 8 Peer Instruction sessions, each of which is between 1.5 and 2 h long. During each session students answer 8–12 ConcepTests on difficult concepts selected from the preclass reading assignment. ConcepTests are short conceptual questions that focus on a single topic [23]. Students answer individually initially and, after discussing each question with their team, they answer again. These sessions occur at the beginning of the class as they allow the instructor to probe students' understanding of the reading and address difficult concepts.

Tutorials: worksheets that are designed to address common misconceptions about the course content. Depending on the semester, we use 6–8 tutorials from the "Tutorials in Introductory Physics" [24] developed by the Physics Education Group at the University of Washington. During this activity, which lasts 1 h, students work with their team on the worksheet and this allows them to explore their thinking about the more difficult concepts in the material.

Estimation activity: students use their physics knowledge and reasoning skills to estimate five quantities, related to the content of the class. Students are given 30 min to think and work with their team to estimate the quantities to the nearest order of magnitude.

Experimental design activity: hands- on, lablike activities or online simulations (typically PhET) [25] that help students develop experimental and analytical skills that are important for the projects.

Problem set reflection: problem sets are comprised of 4–5 physics problems that students are given a week to work on at home. Students are instructed to give the problems their best effort (without consulting others) before coming to class and to bring their solutions to class

to work through with their team, during the problem set reflection activity. During this time students work with their teams to discuss and improve their solutions, resolve conceptual difficulties, and reflect on areas that need to be reviewed. At the end of this activity, students submit their revised solutions with a written reflection on the aspects of the problem set they struggled with and an explanation of how they resolved any misunderstandings.

Readiness assurance activity (RAA): RAAs are assessments conducted in-class, 5 times over the course of the semester, to ensure that everyone is on track in the learning of the basic concepts. During the first half of each RAA, students work individually to answers a set of problems. Students are free to consult the textbook or the internet but are not allowed to discuss the problems with one another. During the second half of each RAA, students get together with their team and discuss the same problems, until they agree on the answer. After reaching a consensus, the team submits their answer as a group. After submitting their team round response, the system indicates whether the response is correct or incorrect. If the response is incorrect, the team has the chance to answer again, for half credit. If the response is incorrect a second time, the team can answer again for quarter credit. If the response is incorrect a third time, the system reveals the correct solution. This second part of the RAA provides an opportunity to learn collaboratively in teams as well as receive immediate feedback. The overall RAA score is determined by a combination of the student's individual score (50%) and their team's score (50%).

D. Online annotation system

As AP50 has no lectures, the reading is the students' primary exposure to the class content. Before each class, students are required to read and annotate a specific chapter in the textbook (see Supplemental Material [22] for spring reading schedule). Students access the textbook via an online collaborative textbook annotation tool [26]. The discussion functionality of the annotation tool is integrated directly into the margins of the (online) course textbook. The annotation tool supports threaded discussions of the material; placement in the margins simply improves the organization of the discussions in context of the textbook material. The textbook is uploaded to the annotation tool website, to which students log on to read and annotate the text. Annotations are made by highlighting a passage of the textbook and typing into a text field that appears in the margin. Students annotate by asking questions about what they are reading and also by responding to other questions and comments made by their classmates. Annotations are organized into "threads," which constitute a starting comment or question followed by all the replies made by other students to the initial annotation or to the subsequent replies. In this way, students have a discussion about specific aspects of the content, within the context of the textbook. In AP50 NB is the only mechanism for delivering the course content to the students, entirely supplanting traditional lectures. It is therefore very important that students read thoughtfully before coming to class, to be prepared to participate in and learn from the in-class activities and from each other.

E. Annotation assessment

Given the important role that the reading assignments play in the structure of AP50, students are assessed on the thoughtfulness of their annotations and this assessment is important in determining each student's final grade. Each class session pertains to a single chapter in the textbook, which students are expected to read before coming to class. Each semester is divided into five units, each of which consists of 3–4 chapters of the textbook. To encourage students to stay on top of the reading, their annotations are evaluated at the end of each unit and this evaluation represents 15% of the overall course grade. To assess the extent to which students' annotations were meaningful and exhibited thoughtful reading of the text, we scored each annotation based on a rubric with a 3-point "quality" scale. Annotations lacking meaningful physics receive a quality score of 0 while factual, definition type annotations receive a quality score of 1. Annotations that justify questions or explanations with substantiated physics concepts receive the maximum quality score of 2. The rubric used to evaluate the quality of annotations is outlined in Table I. Figure 1 provides examples of specific annotations, how they were scored as well as the rationale for the scoring.

III. METHOD

A. Course learning metrics

To study the relationship between students' NB participation and their learning, we use a number of metrics. We measure learning with conceptual surveys, ConcepTests, and exam scores.

1. Conceptual surveys

At the beginning and end of each semester, we administered a conceptual survey as a pre- and post-test; the Force Concept Inventory (FCI) [27] in the fall semester, and the Conceptual Survey on Electricity and Magnetism (CSEM) [28] in the spring. We found that both surveys had high

TABLE I. Rubric for evaluating the quality of NB annotations.

Score	Description or criteria
0	Does not demonstrate any thoughtful reading of the text
1	Demonstrates reading, but no (or only superficial) interpretation of the text
2	Demonstrates thorough and thoughtful reading AND insightful interpretation of the text

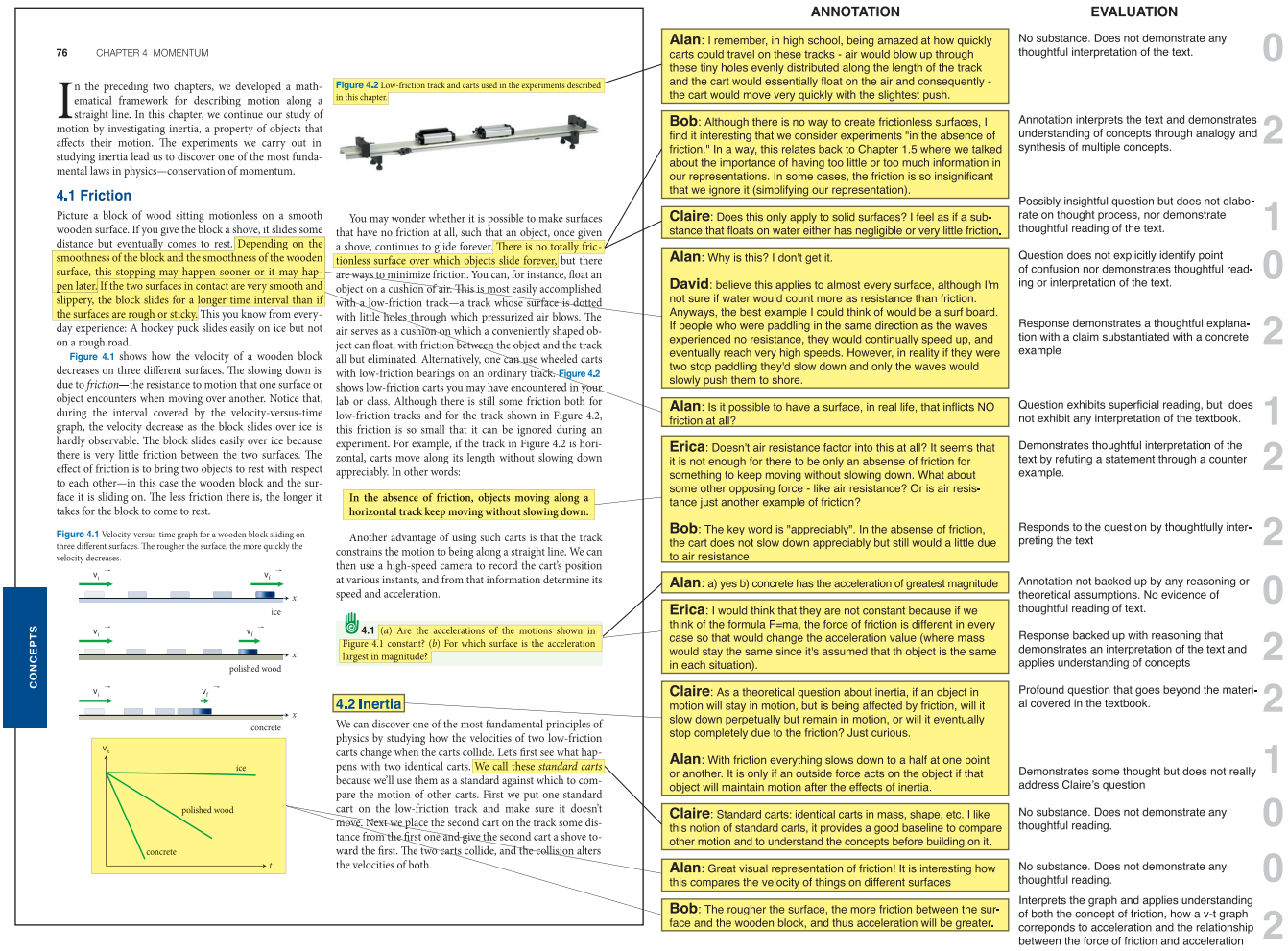

FIG. 1. Examples of annotations, scores, and justifications for scoring.

Kuder-Richardson reliability coefficient values (KR-20), which suggests these surveys have strong internal consistency [29]. The KR-20 of the Force Concept Inventory was equal to 0.90 for the pretest and 0.88 for the post-test. The KR-20 of the Conceptual Survey on Electricity and Magnetism was equal to 0.83 for the pretest and 0.85 for the post-test. These values are comparable to those cited in the literature for the Mechanics Diagnostic Test, which is a conceptual survey that is very similar to the Force Concept Inventory. The KR-20 for the Mechanics Diagnostic Test was found to equal 0.86 for the pretest and 0.89 for the post-test [30]. The range for KR-20 is between 0 (indicative of no internal consistency) and 1 (perfect internal consistency). A KR-20 value greater than 0.80 satisfies the conditions required for comparison of scores between individuals [31]. We found the internal consistency for all the conceptual surveys used in this study met this criterion.

Researchers and practitioners routinely use the normalized gain [32] to evaluate the effectiveness of instruction. Normalized gain is defined as the ratio of the difference between the post and pretest scores to the possible increase in score (post-pre and max score-pre). Normalized gain is a useful way of measuring learning gains of students with different pretest scores [32]. We use students' pre- and postsemester conceptual survey scores to calculate their normalized gain as a measure of how much physics they have learned over the course of the semester. The distributions of the normalized conceptual survey scores (pre and post) as well as the normalized gain for both the spring and fall semesters can be found in the Supplemental Material [22]. Using a two-sample t test, we determined that the average CSEM pretest score in the spring course for students who also took the fall course (11.1 out of 32) is not statistically different from the average CSEM pretest score for students who did not take the fall course (11.6 out of 32) ($p = 0.76$).

2. *ConcepTest performance*

In addition to using students' scores to these pre- and post-tests, we collected their individual responses to ConcepTests posed during the Peer Instruction sessions

during each semester. During the fall and spring semesters the students answered 48 and 54 ConcepTests, respectively. We determined the fraction of ConcepTests that each student answered correctly individually, before peer discussion. Using a two-sample t test, we determined that the average fraction of correct ConcepTest scores in the spring course for students who also took the fall course (0.46) is not statistically different from the average fraction of correct ConcepTest scores for students who did not take the fall course (0.45) ($p = 0.76$). The distribution of these scores for each semester can be found in the Supplemental Material [22].

3. Exam performance

As another metric of student performance, we used students' average performance on five unit exams administered at the end of each unit (approximately every three weeks). These exams are two-part, collaborative exams during which the students complete the questions individually first and then as a group. We use the average of their scores on the individual portion of these exams as another measure of performance in the course. Using a two-sample t test, we determined that the average individual exam scores in the spring course for students who also took the fall course (0.48) is not statistically different from the average individual exam scores for students who did not take the fall course (0.49) ($p = 0.76$). The distributions of average, normalized exam scores for each semester can be found in the Supplemental Material [22].

B. Coding annotations

To measure students' level of engagement in the online discussion we coded the annotations on a number of dimensions. We assessed the thoughtfulness of the annotations using the quality rubric already described (Table I). Figure 2 shows the distributions of annotation quality scores for students in each of the two semesters. As students' average annotation quality score was factored into their final grade, all students' annotations were assessed ($N = 17\,578$) and assigned a quality score. Two physics professors and one physics Ph.D. student performed all quality coding. Before coding independently, the coders conducted two rounds of calibration (in batches of 75 annotations at a time) to ensure internal consistency. The average interrater reliability was 81% for coding annotation quality (kappa = 0.76). These statistics were obtained by averaging the percent agreements between each of three combinations of reviewers (for the 75 annotations all three coders coded after the two calibration rounds).

On average students wrote, and were evaluated on, 110 annotations over the course of each semester. The distribution of the number of annotations that were scored per student, per semester can found in the Supplemental Material [22]. Using a two-sample t test, we determined that the average quality of annotations in the spring course

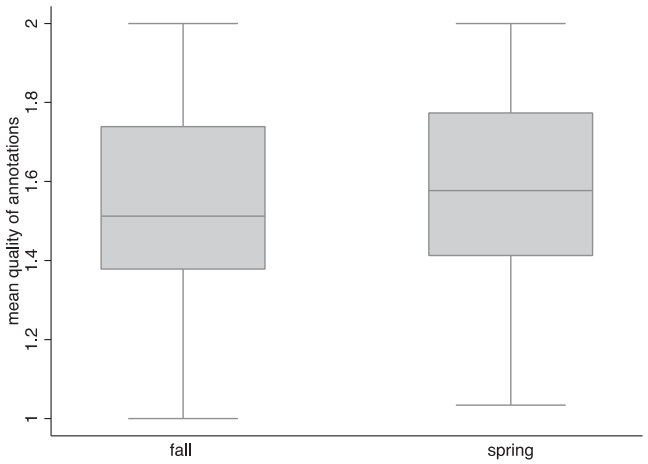

FIG. 2. Box and whisker plots depicting the distributions of annotation quality scores for students in each of the two semesters. The ends of the box are the upper and lower quartiles; the median is marked by the vertical line inside the box; the whiskers are the two lines outside the box that extend to the highest and lowest observations and, therefore, show the range.

written by students who also took the fall course (1.59, $N = 57$) is not statistically different from the average quality of annotations who did not take the fall course (1.56, $N = 14$) ($p = 0.67$).

All annotations were also classified as one of three types: comment, question, or explanation. Comments are typically the first annotation in a thread and are statements about the textbook that are made without the expectation of a reply. Questions, on the other hand, are posed with the expectation of a response with information or an explanation of some concept. Questions can either be the first annotation in a thread or not. Explanations are always written in response to a question and are therefore never the first annotation in a thread. Using these rules, a physics Ph.D. student classified 200 annotations as one of the three types (comment, question, or explanation). The rules were then converted into a code in STATA that automatically classified annotations as comment, question, or explanation. The code was used to classify the 200 human sorted annotations, and agreement between the human and the computer categories was found to be 94% (kappa of 0.85). The remaining 17 378 annotations were classified for type with the code. Figure 3 shows the distribution of comments, questions, and explanations for all the annotations over both semesters ($N = 17\,578$).

C. Coding threads

We coded a subset of discussion threads from the spring semester course using an adapted scheme developed to examine discourse patterns and collaborative scientific reasoning in peer discussions [33]. We define a thread as being made up of at least two annotations (i.e., we ignore isolated annotations). We categorized threads as one of the

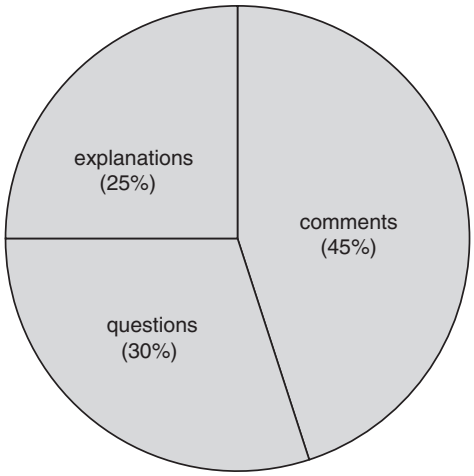

FIG. 3. Distribution of annotation types.

following types: consensual, responsive, transfer, generative, and argumentative. Table II provides examples of each of these five different types of threads. Two physics professors and one physics Ph.D. student coded 6 chapters worth of data, which consisted of 1240 annotations organized into 446 threads. Before coding independently, the coders conducted two rounds of calibration (in batches of 75 threads) to ensure internal consistency. The interrater reliability was 83% (kappa = 0.70). The majority of the thread coding discrepancies were between the elaborative transfer and elaborative generative categories. The percentages in the first column of Table II indicate the proportion of coded threads that fell into each of the five categories. The generative and argumentative threads are of particular interest to us, as these are the types of activities that research has shown to be most effective in promoting

TABLE II. Examples of threads group into the five different categories.

Thread type	Thread conservation
Consensual (5%)	Student 1: Is this talking about an electrically neutral object interacting with the charged tape? Meaning an opposite charge is somehow induced in the neutral object?
	Student 2: Yes! This is explained a little later in the chapter.
Responsive (37%)	Student 1: Obviously, the word "static" in static electricity is the same use of static as having static-y hair for instance, but does it have a connection to the definition so static meaning motionless? If so, what is the connection?
	Student 2: Yes actually. The connection is that the charge just stays there at the surface of the object until it gets a chance to escape via grounding or to an object with smaller or opposite charge. Thus, it contrasts current electricity, which isn't motionless but rather flows through conductors such as wires in energy transmission.
Elaborative transfer (15.5%)	Student 1: Hm. In this case, it's strange because you would think this contributes to flux since it contributes to the number of inward field lines crossing the surface—or is that assuming that it stops within the surface?
	Student 2: Because the charge is outside the surface, it doesn't matter. It cancels itself out.
	Student 3: You are right student 2, because the field line does not stop inside the surface any field line that enters must also leave.
Elaborative generative (15%)	Student 1: What if there is only one charge present? Do the lines then originate or end at infinity?
	Student 2: I think theoretically speaking they do, but as you get farther and farther from the charged particle, the strength of the force decreases (Coulomb's law). Although the field is still in existence at very far distances, or at "infinity" if you will, there really isn't a tangible effect because the particles are so far apart.
	Student 3: What if the lone particle has a negative charge? Since electric field lines emanate from positive to negative, would there be any lines if there was only one negatively charged particle?
	Student 4: I think there would still be field lines pointing toward the negative particle even if it were theoretically the only charged particle in all of the Universe. In reality, humans have never observed any scenario where charges are not balanced in the Universe. Emily's comment on the positive field at infinity is applicable to the negative charge as well.
Elaborative Argumentative (6.5%)	Student 1: Could this also be a rectangle?
	Student 2: Yes, any prism should work. I think they pick a cylinder because there really is only three distinct surfaces and only one variable (radius) matters.
	Student 3: No, maybe not. Making the surface a rectangle instead of a plane would increase the surface area and increasing the surface area decreases the density, it may be more accurate to use a plane instead of a rectangle.
	Student 4: I think the important point is the area of the top surface of the cylinder does not change. Even as the height of the cylinder is increased, the surface area of the top does not change, and neither does the field line density. Thus, we know that the magnitude of the electric field due to the sheet does not change. So, i think you could really use any shape whose side parallel to the sheet does not change in volume as you adjust the distance from the sheet.

TABLE III. Framework used to code NB threads.

Thread Type	Description
Consensual	Only one student contributes substantive statements. Other student responds by passive agreement.
Responsive	Both questions and responses of at least two students contribute to a substantive discussion.
Transfer	Knowledge is shared in what is typically a longer discussion. No new ideas emerge.
Generative	New ideas are linked to someone else's idea and knowledge emerges in a constructivist manner.
Argumentative	Critical discussion during which there is disagreement between participants.

learning [34]. A brief description of each of these categories is summarized in Table III. Figure 4 shows a distribution of the different kinds of threads.

D. Sectioning

To explore the effect of sectioning, we subdivided the class into smaller and smaller online sections, 3 times over the course of each semester. We subdivided the class from the initial size of around 80 students to sections of 40 students, then 20 students, and finally to sections of 10 students. To disentangle the possible effect of the length of time in the course from that of the size of the section, we compared the scores of the 57 students who continued from the fall to the spring course to the 13 students who started the course new in the spring. These two groups of students had been exposed to the course for different periods of time and yet there was no statistically significant difference in their annotations. It is important to note that these sections existed only in the context of the online annotation forum and did not affect the amount of time students spent in class. To create the sections, we first randomly assigned students to sections and then made some adjustments to ensure that each section had the same average on the presemester conceptual survey. This subdivision effectively decreases the size of the online discussion. Instead of all students being able to see and respond to all annotations classwide, they can only see and respond to the annotations from the students in their section. Our goal is to determine the "sweet spot" for the number of students in a particular section. When the discussion group is too large, students will find they have nothing left to say, while if it is too small, interesting discussion topics may be missed. We, therefore, compare the average thread length and average annotation quality as a function of section size to determine if the quality of the discussion is related to the size of the section.

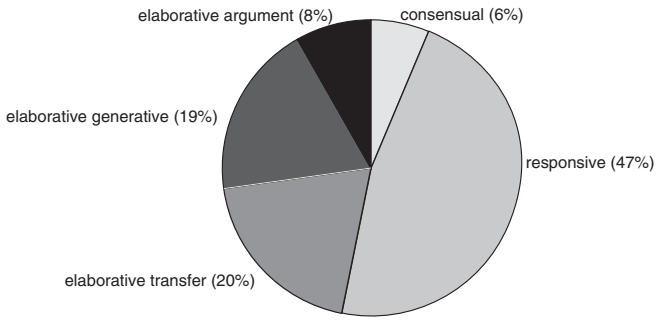

FIG. 4. Distribution of thread types.

E. Seeding

To study the effect of seeding the online discussion we created fictitious student accounts and used these accounts to post high-quality annotations from the previous year's discussion ahead of the class annotating the text. For each of the five units, we seeded half of the sections, leaving the other half unseeded. We changed which sections were seeded and which sections were unseeded from one unit to another. In this way, all students were in seeded sections and all the students were in unseeded sections for different units during the semester. We compare students' annotations (quality, type of thread) when they are in seeded sections to when they are in unseeded sections. For each chapter in the seeded sections, we randomly selected 10 high quality annotations (scored as a 2 according to the quality rubric described in Table I) that had started a thread in the previous year. We imported each seeded annotation into the discussion anonymously. It is common practice for real students in the class to post annotations that are anonymous to everyone (except the instructor). Therefore, to the students, these seeded annotations appear to be no different from "real" annotations.

IV. RESULTS

A. Relationship between students' participation in the discussion and course performance

Table IV shows student correlations (number of students, $N = 161$) between course learning metrics (ConcepTest and exam performance) and five annotation metrics: average quality, number of annotations per semester, ratio of explanations to annotations, ratio of comments to annotations, and ratio of questions to annotations. We use the ratio of type of annotation to total number of annotations so that we can disentangle how many annotations a student submits from whether the student is fundamentally a questioner or a responder. The average annotation quality metric is obtained by finding the average quality score (on the three point scale described in Table I) over all of each student's annotations. The number of annotations per semester metric is raw count of annotations

TABLE IV. Correlations between students' in-class learning metrics and their annotation metrics (number of students, $N = 161$).

	CT performance	Exam performance	Normalized gain
Average annotation quality	0.17[a]	0.31[c]	0.21[a]
Total number of annotations	0.03	0.00	0.21[a]
Ratio of explanations to annotations	0.43[b]	0.36[c]	0.16
Ratio of comments to annotations	0.03	−0.06	−0.06
Ratio of questions to annotations	−0.39[b]	−0.35[c]	−0.11

[a] $p < 0.05$
[b] $p < 0.01$
[c] $p < 0.001$

a student submits over the course of the entire semester. The ratio of comments, explanations, or questions to annotations metric is the fraction of each student's total annotations that are classified as comments, questions, and explanations. Table IV shows that students who write high quality annotations do better on in-class exams and on in-class ConcepTests than students who write low quality annotations. Table IV also shows that students with a high ratio of explanations to annotations (and those who have a low ratio of questions to annotations) perform better on both ConcepTests and exams than students with a low ratio of explanations to annotations. We also find that online engagement (as measured by both annotation quality and the total number of annotations) is correlated with normalized gain. Students who write, on average, higher quality annotations and who participate more in the online forum (by writing more annotations) have higher normalized gains than students who participate less and write lower quality annotations. Table V shows standardized coefficients for a linear regression model predicting students' average exam scores using students' average annotation quality score and ratio of explanations to annotations score as predictor variables and controlling for preclass physics knowledge (precourse FCI or CSEM). Students who engage in high-level discussion online, especially by providing answers to the questions of their peers, perform better on in-class exams than students who do not, even when we control for how much physics students know at the beginning of the semester. This model shows that, controlling for precourse conceptual survey score, a 1 standard deviation increase in average annotation quality predicts a 23% increase in a student's average exam score and a 1 standard deviation increase in the ratio of explanations to annotations predicts a 13% increase in a student's average exam score.

TABLE V. Standardized coefficients for linear regression models predicting average exam performance using students' average annotation quality score and ratio of explanations to annotations score as predictor variables and controlling for preclass physics knowledge (precourse FCI or CSEM). (Number of students, $N = 161$).

	Average exam performance
Average annotation quality	0.23[b]
Ratio of explanations to annotations	0.13[a]
Precourse FCI or CSEM	0.56[b]
R^2	0.48
Root mean square error	0.73

[a] $p < 0.01$
[b] $p < 0.001$

B. Seeding

We find that we can steer online interaction towards more productive and engaging discussion by seeding the discussion. Seeded sections produce longer threads, higher quality annotations, and a greater proportion of generative threads than unseeded sections. We used a series of two-sample, equal variance t tests to determine the difference between seeded and unseeded sections and threads (after determining each population was normally distributed with the same variances). We found a statistically significant difference in the average thread length of seeded threads compared to unseeded threads. Unseeded threads have, on average 0.46 replies, while seeded threads receive an average of 1.16 replies ($p < 0.001$, effect size 0.30). Annotations in seeded sections are made, on average, significantly earlier than annotations in unseeded sections. Annotations in seeded sections are made, on average, 21 h before class whereas annotations in unseeded sections are made, on average, 11 h before class ($p < 0.001$, effect size 0.07). Figure 5 shows that the average quality of annotations in seeded sections exceeds the average quality of annotations in the unseeded sections ($p < 0.05$, effect

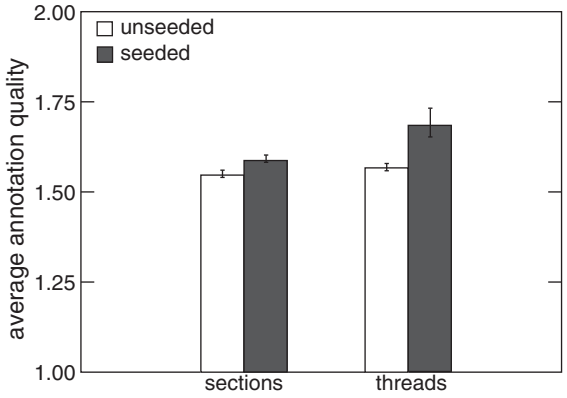

FIG. 5. Average quality of annotations in unseeded versus seeded sections (left) and in unseeded versus seeded threads (right).

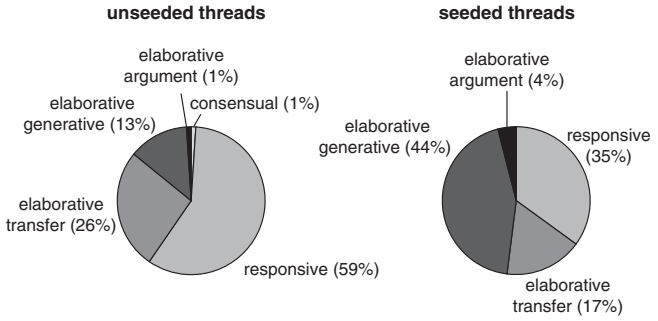

FIG. 6. Percentage of thread types in unseeded versus seeded sections.

size = 0.02). This figure also shows that the average quality of annotations in the seeded threads is significantly higher than the average quality of annotations in the unseeded threads ($p < 0.001$, effect size = 0.13). We also find that seeded threads demonstrate an above average amount of "generative" discussion. Figure 6 shows the fraction of threads that fall into each of the five thread types (described in Table III) for seeded versus unseeded threads. We used an ANOVA analysis of variance to determine that the difference between groups is statistically significant ($p < 0.05$). Especially noteworthy is the large fraction of elaborative generative discussions that emerge in the seeded threads compared to the unseeded threads.

C. Sectioning

We computed both the average thread length and average annotation quality in each of the sections as a function of the size of the section. Figure 7 shows the average thread length increases as the size of the section increases. The correlation between section size and average thread length is 0.74 ($p < 0.001$). Figure 7 also shows the average annotation quality increases as the size of the section increases up to a point ($N = 40$) and then decreases. The correlation between section size and annotation quality is −0.28 ($p < 0.05$). We find that larger sections result in longer discussion threads. However, section size and annotation quality are not linearly related.

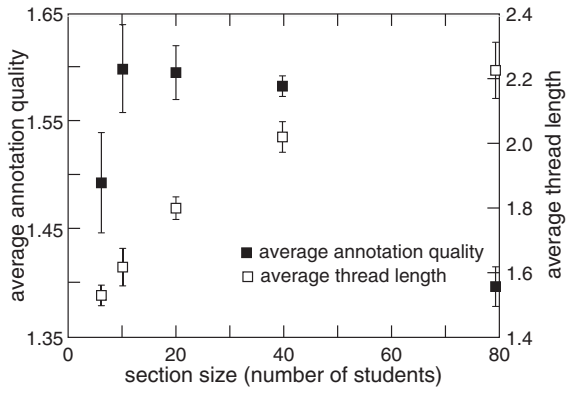

FIG. 7. Average thread length and average annotation quality as a function of section size.

V. DISCUSSION

A. Relationship between students' participation in the discussion and course performance

We find that students who write high-quality annotations perform better on exams and the postsemester conceptual survey than students who write low-quality annotations. We also find that students who annotate more with explanations tend to do better on in-class activities (exams and ConcepTests) than students who annotate mostly by asking questions and making comments. The positive relationship between meaningful engagement in online discussions and academic performance supports the findings of earlier studies [9,11,17]. The fact that these relationships are statistically significant even after controlling for students' physics knowledge at the beginning of the course (pre-semester FCI and CSEM scores) is important. The relationship between annotation quality and students' performance on these two activities is independent of how much physics a student knows at the beginning of the semester. This relationship is perhaps not surprising given that our rubric for measuring annotation quality is, to a certain extent, measuring the amount of effort students put into annotating thoughtfully. Despite this, we did not find a relationship between any student learning metric and any other annotation metric one would associate with effort. There is no correlation between students' performance in class and the number of annotations they make or the amount of time they spend annotating (except for a small correlation between number of annotations and normalized gain). Based on this, it seems as though the rubric we use to assess thoughtful reading is not only measuring the amount of effort each student puts into annotating the reading but is also measuring the degree to which they think about and synthesize the material. Simply putting more time into annotating is not associated with increased performance in learning metrics. Students who put more effort into producing *thoughtful* annotations (regardless of their physics ability) do better on exams and the post-semester conceptual survey than students who put in less effort. The most significant correlations are those between student exam performance and annotation quality and between exam performance and the ratio of explanations to annotations. There is also a strong (negative) correlation between exam performance and the ratio of questions to annotations which is, at least in part, due to the fact that students who annotate with a higher proportion of explanations, by definition, annotate with fewer questions. The fact that there is zero correlation between the number of annotations and exam performance is interesting because it suggests that the quality of the interactions is more predictive of higher performance than the quantity of the

interaction. We find a weak, although statistically significant relationship between the normalized gain and both the average annotation quality and total number of annotations. This indicates that students who write high-level annotations *and* students who annotate more have higher normalized gain than students who write lower-level annotations and students who annotate less. The fact that annotation quality is less predictive of normalized gain than exam performance makes sense given the nature of the performance metric. The conceptual surveys (used to measure normalized gain) only measure the mastery of a subset of topics. The Force Concept Inventory, for example, only measures students' understanding of Newton's laws. The in-class exams, on the other hand, are designed to test the students understanding of a broader range of topics emphasized in the course and are therefore a better measurement of students' mastery of the preclass reading.

B. Seeding

We show that it is possible to seed prior-semester comments into the new semester's discussion to stimulate an above-average amount of discussion. Additionally, this discussion demonstrates an above-average amount of "generative interaction," the interaction type demonstrated to be of greatest value for learning [34]. We have also found that students in seeded sections annotate and read significantly earlier and their annotations are of better quality compared to students in unseeded sections. The fact that, by seeding the discussion, instructors can influence students' annotating behavior is very important given the relationship we have shown between students' effort in thoughtful discussion online and their conceptual learning over the course of the semester. By seeding the discussion, instructors can increase the amount of thoughtful effort students put into their online participation. This finding is contrary to another study that looked at the relationship between the level of thinking discussion prompts and the related responses [35]. In this study, each student was required to post one prompt for discussion during the course of the semester. The level of thinking for both the prompts and subsequent responses were evaluated using Bloom's taxonomy. There was no relationship found between the level of thinking in the prompt and that in the response [35]. Our finding that seeding insightful prompts can effect the quality of responses in the discussion has powerful implications for instructors of flipped classrooms who are interested in increasing the likelihood of students reading the material before coming to class and engaging in higher quality discussion.

C. Sectioning

As the size of the section increases, students initiate threads less and instead add on to existing threads. The correlation between the number of initiated threads per student and section size is -0.30 ($p < 0.05$) and the correlation between the number of replies per student and section size is 0.56 ($p < 0.001$). These findings lend support to the hypothesis that when there are too many participants in a discussion, it becomes saturated with annotations and there is nothing left to say. In larger sections, students might be adding comments to existing threads rather than starting their own threads due to this saturation effect. It remains to be determined whether it is beneficial for "saturation" to force students to reply to threads rather than initiating their own. As we have argued, conversation has been recognized as having higher impact for learning, which suggests it could be beneficial to force students into conversation. However, if those forced conversations are filled with "me too" statements reflecting the student's requirement to comment, there may be no beneficial dialogue. To provide a better sense of the relationship between section size and the effectiveness of the online discussion, we also studied the relationship between the number of students in a section and the average annotation quality in that section. The average quality of annotations increases as the section size increases, but only up to a certain section size. Beyond a size of 40 students, the average quality of the annotations decreases, which suggests that the saturation effect is problematic for sections exceeding this size. Because of the constraints of our study, we were not able to collect data on sections with more than 20 students but less than 40. Further research is necessary to fill this gap as well as to study the effect of saturated discussion forums on student learning. Do students learn more in online discussions when they are forced to reply to threads than when they can initiate their own threads? Another interesting area for future research would be to determine the relationship between when, relative to class, students participate in the online discussion and their academic performance. Understanding the relationship between reading and engaging in an online discussion before class and students' performance in class would provide useful insight into the structure of flipped classroom environments.

VI. LIMITATIONS

The fact that we do not have a control group of students who took the course but did not participate in the online reading forum means that we cannot draw causal inferences about the effect of participation in online reading discussions on student learning. We also cannot completely disentangle the amount of effort a student puts into annotations from the quality of those discussions. We are in the process of planning future studies to address these limitations as well as to better understand the relationship between section size and annotation quality.

VII. CONCLUSION

We have shown that online reading discussions are an effective mechanism to engage students in a flipped

classroom. We have also found a positive relationship between students' conceptual learning and the quality of their participation in the online discussion. Given this finding, instructors should incentivize high-quality interactions in the context of online reading. By seeding discussions with thought-provoking content and managing section size, we have also gained insight into how students' online conversations can be managed and guided to promote learning. Seeding discussions with successful threads leads to more constructive discussions. Limiting the size of the online discussion to 20–40 students optimizes both the quality and length of the discussion.

ACKNOWLEDGMENTS

Several people contributed to the work described in this paper. E. M., D. K., and K. M. conceived of the basic idea for this work. K. M., D. K., S. Z., and J. Y. designed and carried out the experiments, and analyzed the results. E. M. and D. K. supervised the research and the development of the manuscript. K. M. wrote the first draft of the manuscript; all authors subsequently took part in the revision process and approved the final copy of the manuscript. Marinna Madrid and Michael Mobius provided feedback on the manuscript throughout its development.

[1] C. C. Bonwell and J. A. Eison, Active learning: Creating excitement in the classroom, ASHE-ERIC Higher Educ. Rep. **808**, 19 (1991).

[2] A. W. Chickering, Z. F. Gamson, and M. Sorcinelli, Research Findings on the Seven Principles, *Applying the Seven Principles for Good Practice in Undergraduate Education, New Directions in Teaching and Learning* (Jossey-Bass/Wiley, San Francisco, 1987).

[3] P. A. Rovai, Building sense of community at a distance, Int. Rev. Res. Open Dist. Learn. **3**, 1 (2002).

[4] J. Bradshaw and L. Hinton, Benefits of an online discussion list in a traditional distance education course, Turk. Online J. Dist. Educ. **5**, 1 (2004).

[5] R. T. Berner, The benefits of bulletin board discussion in a literature of journalism course, The Technology Source Archives at the University of North Carolina; http://technologysource.org/article/benefits_of_bulletin_board_discussion_in_a_literature_of_journalism_course/ (2003).

[6] M. Tallent-Runnels, J. Thomas, W. Lan, S. Cooper, T. Ahern, S. Shaw, and X. Liu, Teaching courses online: A review of the research, Rev. Educ. Res. **76**, 93 (2006).

[7] A. L. Brown and J. C. Campione, *Psychological Theory and the Design of Innovative Learning Environments: On Procedures, Principles, and Systems* (Lawrence Erlbaum Associates Inc. Hillsdale, NJ, 1996).

[8] D. Nandi, S. Chang, and S. Balbo, A conceptual framework for assessing interaction quality in online discussion forums. Same Places, Different Spaces, *Proc. Ascilite Auckland* (2009).

[9] S. B. Smith, S. J. Smith, and R. Boone, Increasing access to teacher preparation: The effectiveness of traditional instructional methods in an online learning environment, J. Spec. Educ. Techn. **15**, 37 (2000).

[10] M. Durham, Computer conferencing: Students' rhetorical stance and the demands of academic discourse, J. Comput. Assist. Learn. **6**, 265 (1990).

[11] S. L. Althaus, Computer-mediated communication in the university classroom: An experiment with on-line discussions, in *Communication Education* (Taylor & Francis, London, 1997), pp. 158–174.

[12] M. Durham, Computer conferencing, students' rhetorical stance and the demands ofacademic discourse, J. Comput. Assist. Learn. **6**, 265 (1990).

[13] P. Gorski, R. Heidlebach, B. Howe, M. Jackson, and S. Tell, Forging communities for educational change with e-mail discussion groups, Multicultural Perspectives **2**, 37 (2000).

[14] S. M. Zvacek, What's my grade? Assessing learning progress, TechTrends : for leaders in education and training **43**, 39 (1999).

[15] M. Scardamalia, C. Bereiter, R. S. McLean, J. Swallow, and E. Woodruff, Computer-supported intentional learning environments, J. Educ. Comput. Res. **5**, 1 (1991).

[16] D. C. Edelson and K. O'Neill, The CoVis collaboratory notebook: Computer support for scientific inquiry, *Proceedings of the Annual Meeting of the American Educational Research Association, New Orleans, Lousiana* (New Orleans, 1994).

[17] V. Light, Let's you and me have a little discussion: Computer mediated communication in support of campus-based university courses, Studies High. Educ. **25**, 1 (2000).

[18] L. Romeo, Asynchronous environment for teaching and learning: Literacy trends and issues online, The Delta Kappa Gamma Bulletin **6**, 3 (2001).

[19] J. Sheard, S. Ramakrishnan, and J. Miller, Modeling learner and educator interactions in an electronic learning community, Australas. J. Educ. Tech. **19**, 2 (2003).

[20] P. C. Blumenfeld, E. Soloway, R. Marx, J. Krajcik, M. Guzdial, and A. Palincsar, Motivating project-based learning: Sustaining the doing, supporting the learning, Educ. Psychol. **26**, 369 (1991).

[21] L. K. Michaelsen, A. B. Knight, and L. D. Fink, *Team-based Learning: A Transformative Use of Small Groups* (Greenwood Publishing Group, Westport, 2002).

[22] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.12.020143 for course schedule and descriptive statistics showing the distributions of annotation and course performance metrics.

[23] E. Mazur, *Peer Instruction A User's Manual* (Prentice-Hall, Upper Saddle River, NJ, 1997).

[24] L. C. McDermott and P. S. Shaffer, PER (1998). Tutorials in introductory physics (Prentice Hall, Inc., NJ, USA, 2001).
[25] PhET Simulations. https://phet.colorado.edu/en/simulations/category.
[26] NB Annotation System nb.mit.edu.
[27] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[28] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).
[29] R. P. McDonald, *Test Theory: A Unified Treatment* (Lawrence Erlbaum Associates, Mahwah, NJ 1999).
[30] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef, The puzzling reliability of the Force Concept Inventory, Am. J. Phys. **79**, 909 (2011).
[31] J. M. Bland and D. G. Altman, Cronbach's alpha, Br. Med. J. **314**, 572 (1997).
[32] R. R. Hake, Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).
[33] K. Hogan, B. K. Nastasi, and M. Pressley, Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions, Cognit. Instr. **17**, 379 (1999).
[34] M. T. H. Chi, Active-constructive-interactive: A conceptual framework for differentiating learning activities, Top. Cognit. Sci. **1**, 73 (2009).
[35] M. M. Christopher, J. A. Thomas, and M. K. Tallent-Runnels, Raising the bar: Encouraging high level thinking in online discussion forums, Roeper Rev. **26**, 166 (2004).