Structuring Interactions for Large-Scale Synchronous Peer Learning

D Coetzee[‡], Seongtaek Lim[†], Armando Fox[‡], Björn Hartmann[‡], Marti A. Hearst[†‡]
UC Berkeley School of Information[†] and EECS[‡]
{dcoetzee,stlim,fox,bjoern,hearst}@berkeley.edu

ABSTRACT

This research investigates how to introduce synchronous interactive peer learning into an online setting appropriate both for crowdworkers (learning new tasks) and students in massive online courses (learning course material). We present an interaction framework in which groups of learners are formed on demand and then proceed through a sequence of activities that include synchronous group discussion about learner-generated responses. Via controlled experiments with crowdworkers, we show that discussing challenging problems leads to better outcomes than working individually, and incentivizing people to help one another yields still better results. We then show that providing a *mini-lesson* in which workers consider the principles underlying the tested concept and justify their answers leads to further improvements. Combining the mini-lesson with the discussion of the multiple-choice question leads to significant improvements on that question. We also find positive subjective responses to the peer interactions, suggesting that discussions can improve morale in remote work or learning settings.

Author Keywords

Peer Learning, Group Discussion, Crowdwork

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Online learning should have a social component [32]; how to do that in the context of Massive Open Online Courses (MOOCs) is an open question. This research addresses the question of how to improve the online learning experience in contexts in which students do not have the opportunity to interact with the instructor directly and do not know other students. We wish to show how the positive learning results that have been found in in-person classes and Computer-Supported Collaborative Learning (CSCL) contexts can be successfully replicated and extended in a more impersonal online context.

CSCW '15, March 14-18, 2015, Vancouver, BC, Canada.
ACM 978-1-4503-2922-4/15/03.
http://dx.doi.org/10.1145/2675133.2675251

Our approach is inspired by the literature of structured peer learning [40, 51] as well as newer work by Mazur and other large-course instructors who have brought active peer learning into the classroom [10, 50, 12]. The goal of peer learning, also known as collaborative learning, cooperative learning, and in some contexts, peer instruction, is for students to work together in small groups to enhance their own and one another's learning. Peer learning in the physical classroom consists of activities in which students confer in small groups to discuss conceptual questions and to engage in problem-solving. Literally hundreds of research studies and several meta-analyses show the significant pedagogical benefits of peer learning, including improved critical thinking skills, retention of learned information, interest in subject matter, and class morale [4, 25, 34, 17, 41, 40, 51, 50, 12].

Rather than trying out untested designs on real live courses, we have prototyped and tested the approach using a crowdsourcing service, Amazon Mechanical Turk, (MTurk) [28] on a critical thinking task with multiple-choice answers. While crowd workers likely have different motivations from MOOC students, they do commonly encounter training materials when learning new tasks. In addition, the remote individual work setting without peer contact resembles today's MOOC setting where most students learn in isolation.

Our contributions consist of a method for structuring distributed synchronous peer-learning interactions; a software framework that facilitates such interactions and is positively perceived by crowd workers; and a set of experiments demonstrating that these interactions improve crowd workers' performance on a multiple-choice critical reasoning task. We have also successfully deployed these techniques in a large blended course with more than 1000 students. How applicable these results may be to MOOCs are considered in detail in the discussion.

We conducted two sets of crowd worker experiments. The first tested the effects of synchronous small-group discussions, comparing different size groups, and tested the effects of a bonus condition that rewarded cooperation. Motivated by learning sciences findings that students benefit if they have a dedicated time for generating questions about the concepts before a group discussion begins, the second set of experiments tested the effects of providing a "mini-lesson" about the underlying concepts and having workers generate concepts in advance of a discussion about those concepts. The specific findings of the experiments are:

- Small group discussion of provided multiple choice answers improved results over working alone, and
  - Incentivizing people to help one another further improved the results
  - Participating more in discussions was correlated with getting more correct answers.
- Showing people instructional material about the task and asking them to analyze the task in light of that material improves results over no instructional material.
- Discussion of the instructional material exercise is not more effective than generating responses alone to that material.
- Discussion of provided multiple choice answers after analyzing instructional materials improves results additively,
- A majority of participants enjoyed discussing the material.

These experiments inform the creation of prototype technology to bridge the gap between how peer learning is used in traditional classroom settings and how it can be adapted to a MOOC. We find that structured peer discussion produces improved results for workers in learning tasks. In addition, social interaction can be a way to improve the experience of remotely working through learning material. An advantage of this approach is the potential to go beyond what is possible in physical classrooms and determine group formation on the fly based on answers to questions, organizing students according to their current misconceptions about class materials, and pairing them with students most likely to form a good discussion.

BACKGROUND AND RELATED WORK

Empirical Support for Peer Learning

Much convincing evidence has accumulated that if students are asked to discuss the answers **with other students**, their understanding of the material increases more than if they did an active learning component on their own [25, 3], and structured group work can promote problem solving at a higher level than possible with individual effort alone [40]. As explanation, learning theorists often point to Vygotsky's writings about the interdependence between individual and social processes in learning and development [43]. For college students, Astin [1], in summarizing a longitudinal study from 139 degree-granting colleges, explains: First, students may be motivated to expend more effort if they know their work is going to be scrutinized by peers; and second, students may learn course material in greater depth if they are involved in helping teach it to fellow students [41].

Even very simple peer learning methods can be effective. For instance, Ruhl et al. [46] conducted a controlled experiment with 72 undergraduate education students comparing a standard lecture format to the simple "pause procedure" in which students formed dyads and discussed lecture material for 2 minutes three times per 45 minute lecture. Students in the pausing group performed one or two points better on after-lecture quizzes and a course-level comprehension test.

Smith [50] showed that peer discussion enhances understanding even when none of the students in the discussion group originally knows the correct answer. In an in-person introductory genetics course, 350 students answered multiple choice questions – using clickers – throughout the semester. In each case they answered first individually, and then after discussing the question with neighbors, they were allowed to change their answer. When given a second, isomorphically similar problem, despite not having been told the correct answer for the first question, the average scores for the second question of the pair were significantly higher than for the first, and gains were greatest on the hardest problems.

Applying Peer Learning

The simplest form of peer learning is the informal groups [25] or "Quick Thinks" usage [45], in which small groups of students briefly discuss material that has been presented and then decide independently how to respond to some activity, typically answering a multiple-choice question. Substantially greater learning gains have been found when peer discussion is combined with social interaction, for example, facilitated by clickers or similar technology, a result that has been replicated in a number of studies [39, 24].

Successful structured peer learning requires attainment simultaneously of *positive interdependence* and *individual accountability*. Positive interdependence requires interdependent contributions from students in the group; students must also be instructed to emphasize learning over "getting the answer." Thus, instructors must devise activities that ensure a reason for the members of the group to interact with and help one another, and both instructor direction and structure of activities must make social loafing difficult.

Students learn in part by explaining, and those who do not understand can benefit from peers' explanations. To motivate more and better understanding, in-class quizzes can be conducted using structured peer learning by first having every student take the quiz individually and then having the group take the quiz together, with both sets of answers counting toward the grade [55].

The literature provides many examples of how to both succeed and fail at introducing peer learning; many readings describe in detail mechanisms for designing successful group activities [55, 59, 40] and how to overlay other pedagogical techniques on peer learning, including Problem Based Learning (accounting) [9], the Argument-Claim-Evidence Method (Engineering) [49], and the Cognitive Tools and Intellectual Roles [43].

Fischer et al. [14] show that efficient collaborative learning is rarely achieved simply by putting learners together without some supportive instruction. While we focus on peer learning rather than collaborative learning, our work does provide such support in the form of instructional narratives and peer discussion. Fischer also cites specific mechanisms by which group co-construction of knowledge occurs during discussion, such as resolution of socio-cognitive conflicts. While our work does not focus on detecting such activity, the ease of data collection and analysis afforded by our apparatus could be beneficial for collaborative-learning researchers.

CSCL Research on Chat in Online Learning

Peer learning in an online context has been studied in the field of Computer Supported Collaborative Learning (CSCL). As Stahl and others note [52], Vygotsky argued that learning takes place in dyads or groups before it takes place by individuals, underscoring the importance of studying and enabling group cognition.

CSCL research in online collaborative learning has generally focused on asynchronous discussion, such as online discussion forums. Where synchronous interaction has been studied (such as real-time chat), most work has studied online technologies designed to be used in conjunction with in-classroom group learning [5] or has focused on interaction between a pair of students or a student and a tutor, as in [16], rather than a small group of peer learners. Recent work in chat-in-CSCL [54] acknowledges that even though this "paradigm shift" from teacher-communicated learning to discourse-based learning is widely accepted in the CSCL community, there are few theories and applications for analyzing such interaction.

Also related is the literature on online teamwork, which at times intersects with that of online learning, as in for example, Tausczik and Pennebaker [53].

Social Interaction in MOOCs

Stahl and colleagues [35] note that despite years of CSCL literature demonstrating the importance of collaborative learning for online education, neither MOOCs nor Khan Academy offer support for students to interactively explore topics themselves or with peers. Some MOOC promoters likewise feel that while peer instruction is promising, disseminating and adopting it are big challenges even in brick-and-mortar classrooms [38]. Nonetheless, there persist calls for student interaction in MOOCs [32].

Currently the main kind of social interaction supported in MOOCs are online forums; these are usually threaded discussion lists in which comments and questions are visible to everyone in the class. Forums are pervasive in MOOCs and have been characterized as "an essential ingredient of an effective online course" [36], but early investigations of MOOC forums show struggles to retain users over time. In one example, half of forum users ceased participation due to factors such as lack of forum facilitation, an overwhelming number of messages, and rude behavior of other students [36]. As an alternative form of engagement, informal groups associated with MOOC courses tend to spring up around the world. Students use social media to organize groups or decide to meet in a physical location to take the course together.

Great strides have been made in peer grading in MOOCs, where students improve their own understanding by assessing the work of others [29]. However, students in this setting do not interact with one another directly but instead review one another's work anonymously and asynchronously. Work has also been done on self-evaluation of assignments within MOOCs [60], replicating earlier results in smaller class settings showing that self-evaluation helps improve student outcomes [47]. Dow et al. [13] found that self-assessment of crowd work was just as effective as feedback from an expert judge on a consumer review writing task, although expert assessment resulted in more work activity.

Group Work in CrowdSourcing

Little research addresses how to support group work in crowdsourcing platforms. The closest is that of Zhu et al. [62] who studied how to improve productivity and quality of crowd worker output, motivated by concerns about overhead in coordinating workers (also known as process loss [42]) and the potential for groups to be swayed toward an undesirable outcome by factors relating to social influence [31, 23].

Zhu et al. [62] found that asking workers to review the work of others yielded better results on a post-test than having the workers do the work directly without any training. They also found that groups working together semi-synchronously using a shared editing tool performed better on 5 tasks (brainstorming, summarizing a paragraph, writing a product review, solving a moral dilemma, doing a mathematics problem) than individual workers, but that pooling and averaging the results of individual workers improved results still more. The one exception to the pooling result was the mathematics problem, where interactive discussion was necessary for insights to occur and workers to learn from one another. This is the kind of problem that we explore in this work, where group interaction is more likely necessary to arrive at understanding.

APPROACH

Incentivizing Collaboration

In standard peer learning, students are asked to work together to answer a question, either after reading material, hearing a lecture, watching a video, or as part of a homework assignment. Peer learning combines *individual accountability* to ensure that students work on their own with *positive interdependence* to ensure that all students in a learning group participate in group work [25, 40]. In this study, we ask MTurk workers to assume the roles of students and discuss the answers to a question posed to them.

One purpose of these experiments is to see if the design facilitates discussion of problems before introducing it into a MOOC. An important open question is whether, in the absence of the social and academic pressures of a real classroom, workers will engage in substantive discussions. Another reason to experiment is to determine how to design the user interface and answer questions including how much time to allow for individual vs group discussion, how many participants to include in a room, and what kinds of instructions should be given about proper conversational structure.

A potential drawback of using crowd workers is that they may not be motivated to have a productive conversation or try hard to get the correct answer. Therefore, in the first experiment, we manipulated a condition in which if **all** participants converged on the correct answer, they would **all** receive a bonus. This strategy is motivated by a technique from the peer learning literature in which students are incentivized to help the others in their group because their grade on a subsequent quiz will be a combination of the scores of everyone in their group

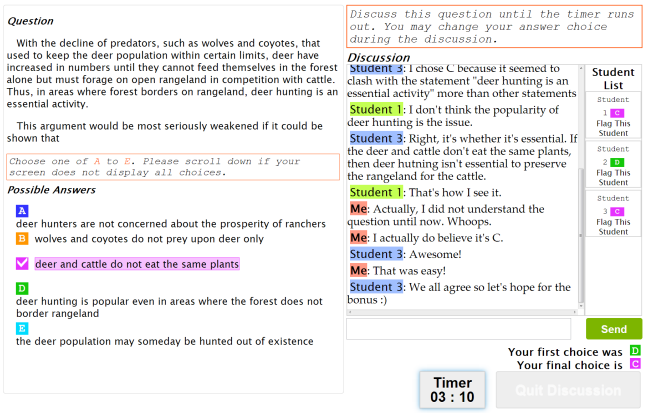

Figure 1. The chat system showed a GMAT practice question on the left while participants discussed choices on the right. In this real example, all participants selected the correct answer and received a bonus payment.

[55]. We hypothesize that this condition would better replicate the situation of earning grades in a course, by motivating the participants to try harder to figure out the right answer, and explain their reasoning to one another, which is a fundamental principle behind peer learning. To facilitate comparison we also compared the effect of offering solo workers a bonus vs. no bonus.

User Interface

We developed a custom web browser-based tool which displays a series of informational screens, many with a countdown timer, the ability to enter text strings and select among multiple choice options, and a chat room for discussion among many students or for individual reflection in the single user condition (see Figure 1).[1]

We decided on text chat rather than video for several reasons:

- Text-based interaction from students is standard in MOOCs and far more reliable than video-conferencing; many MOOC instructors wish to reach students who do not necessarily have access to high-speed internet connections required for real time video conferencing,
- Many students participating in MOOCs are non-native speakers of the instructors' language, and research shows that in some cases people in the non-dominant culture are just as or more active in text-based communication than video or face-to-face [58, 48].
- Research on the relative affordances of text chat, telephone voice communication, and video communication versus face to face communication, including Walther's Social Information Processing Theory of Computer Mediated Communication (CMC), finds that communicators deploy whatever communication cue systems they have at their disposal [56]. Walther's work has found that people can express affect and interpersonal affinity as effectively using an online chat system as in a face-to-face setting, making use of the verbal cues available to them [57]. However, different studies find different outcomes, depending in part on the stated goal (trust, information exchange, etc.).

[1]The code used in these experiments was originally written using Node.js and has since been rewritten in Rails. For Experiment 2 some minor changes were made to how groups were formed using Mechanical Turk, and minor changes were made to the appearance of the user interface as compared to Experiment 1. The source code can be found at **`http://github.com/stlim0730/MoocChat`**

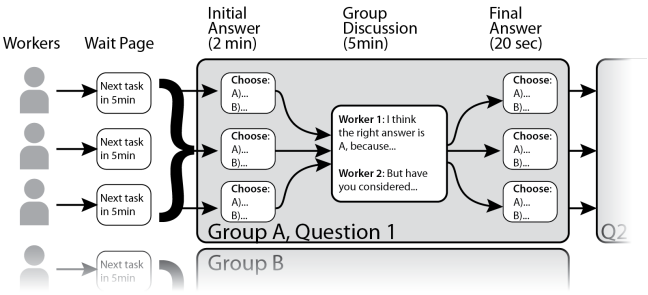

Figure 2. Illustration of worker synchronization for Experiment 1.

Synchronizing Workers

Mechanical Turk does not provide a mechanism to synchronize workers, but we build on earlier investigations that have shown how to assemble multiple crowd workers on online platforms to form synchronous on-demand teams [30, 2]. Our approach was to start tasks (called HITs) at fixed times, announcing them ahead of time and showing a countdown timer until those HITs began. We announced the availability every 5 minutes and had no problem filling HIT sets of size 100, looking for groups of 3. If there was an "odd person out," we placed that worker in their own solo room so their time was not wasted. In some cases, a group of 3 became a group of 2 when a crowd worker who initially accepted a HIT dropped out. This "gather-and-group" mechanism was designed for the MOOC setting, where students may arrive at any point in time; adjusting the waiting time before groups are formed allows us to configure it to suit any arrival rate, and allows students to reliably anticipate when the activity will begin so that they can focus on other tasks.

The method for synchronizing crowd workers for Experiment 1 is illustrated in Figure 2, corresponding to the workflow we call *Discuss Question* (other workflows are described below). While waiting for the timer countdown, participants were shown an information page describing upcoming steps including screenshots of sample questions and the chat room. Once the countdown completed, workers were shown the first question and given instructions to read it and choose one of the answers. The answer could be changed any number of times during this time period, while a timer visibly counted down. Once time ran out, the worker was placed into the chat room, shown the labels for the other workers (Student 1, Student 2, Student 3) and was instructed to begin discussing the question and encouraged to change their answer if appropriate. Again, a timer counted down the time. When there were only 30 seconds left, the screen showed a message in a prominent manner, instructing the worker to make their final choice of an answer. If no final answer was chosen, a null value was indicated in the logs. If more than 45 seconds went by without a message, the system sent a message to the chat room reminding participants to keep the conversation going. In the

manipulated (bonus) condition, the initial message sent to the chat room reminded the workers that if all participants arrive at the same correct answer, they would receive a bonus.

In some cases a single worker was left over after random grouping. These workers went through the same steps and had the same opportunities to revise responses, but instead of engaging with others were prominently asked to explain their response. This controls for benefits due to self-explanation.

Test Materials

We sought questions that could be answered without requiring specialized knowledge, that would likely benefit from careful thought and discussion, and could not be answered by a simple web search. To this end, we obtained a private set of practice questions for the Critical Reasoning Task of the Graduate Management Admissions Test (GMAT). These tasks consist of three parts: a brief passage of text, a question, and 4 or 5 answer choices, of which only one can be selected. The questions are usually related to the construction of arguments, such as finding the evidence to support a conclusion, determining which of a set of statements weakens an argument, or determines which of a set of statements is most strongly supported by the initial paragraph. Figure 3 shows a sample essay and answer choices. Most U.S. business schools use the GMAT as an entrance exam, and more than 250,000 students take the GMAT each year.

Multiple-choice questions and to a lesser extent short-answer questions ("cued recall") are ubiquitous in MOOCs [7, 37]. Multiple-choice questions in particular are used both in quizzes that test comprehension and in assessment instruments for certificates of completion. The "testing effect" finds that frequent testing helps long-term retention [21], and recent results show that multiple choice questions can be effective at both eliciting the testing effect and increasing performance on later short-answer transfer questions [15]. Thus the testing instruments we employ are representative of MOOC tasks and conform to common peer instruction methods.

EXPERIMENT 1

The first experiment assessed if a measurable improvement would occur if workers discussed the problem in groups. It also assessed the effects of varying group sizes and tested whether or not providing an incentive to help other workers would in fact lead to better results. Because the workers do not know the correct answer, if they want to earn the bonus they should have a good incentive to try hard to understand the question well and ensure that even the most certain sounding group member is actually correct in their assumptions.

Participants

Participants were recruited on Amazon's Mechanical Turk with the qualifications of having 95% acceptance rate on 1000 tasks or more. Payment was $2 for the HIT and $2 for each bonus achieved. A given worker could do the HIT only once.

Study Design

Sessions: In a between-participants design 3x2 design, workers either worked alone (solo) or were in placed in a chatroom with a total of 2 or 3 workers and were either offered a bonus or not. Each session consisted of 2 questions and a followup questionnaire which asked for subjective responses to the time allocations and the quality of the discussion. Participants were also asked to optionally provide any other comments or thoughts.

Essay: Tests demonstrating that the olfactory bulbs (relating to the sense of smell) of salmon will react upon exposure only to water from their spawning ground and no other water suggest the possibility that repeated exposure to a specific odorant of their spawning area during the first few weeks of life could stimulate olfactory receptor sites, increasing olfactory sensitivity to that single scent and influencing salmon migration to their spawning area.
Instructions and Possible Answers:
Which if the following, if true, would weaken the conclusion? (answer is E)
(A) Salmon have been trained, through repeated exposure, to recognize phenol and distinguish it from closely related pchlorphenol, in concentrations as small as five parts per billion.
(B) A dog can detect and distinguish the odor of a single human fingerprint for as long as six weeks, after which time the scent usually fades away.
(C) Salmon have been successfully stocked in rivers where special salmon "runs" have been constructed.
(D) Women are acutely aware of the odor of a synthetic steroid named exaltolide, which most men are unable to detect.
(E) Salmon spawned and raised in hatcheries will return to the spawning ground of their parents, and not to the source of the water (the hatcheries) in which they were spawned.

Figure 3. Sample GMAT Critical Reasoning Question (number 4).

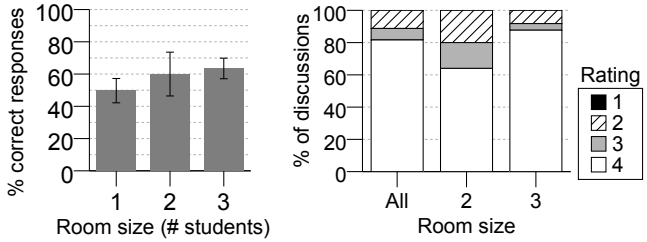

Figure 4. Left: The mean percentage of correct responses is higher in chatrooms with more than one student (Fisher's exact test, $p < 0.01$). Error bars show 95% confidence interval. Right: Most discussions were rated as substantive (4), and no discussions were rated as irrelevant (1).

Timing: We conducted initial tests in which participants were allotted 2 minutes to answer the question individually, 3 minutes for the discussion period, and then 20 seconds to make the final answer choice. After these tests, many participants indicated that they needed more time in the discussion period. The results reported here used the settings of 2 minutes for initial individual answer, 5 minutes for discussion, and 20 seconds for final answer for each question. In discussion, participants without discussion partners were permitted to continue after sending a single message of self-explanation.

Manipulated Conditions: While doing the initial individual answering of the question, participants were shown a message stating that after they chose their initial response, they would join a chat room to discuss their answer with either one or two other people. They were also told at this time if all participants in the chatroom arrived at the same answer and that answer was correct, they would receive a monetary bonus.

	Below, Student 1 initially does not have the correct choice.
Student 3	I hesitated on C
Student 2	I did as well I was trying to decide between a and c
Student 2	what other answer did you consider
Student 3	Well I was still reading and saw the timer at 1 second so I hit C.
Student 3	But
Student 1	I considered B and C
Student 1	B
Student 3	I'm going over iit again
Student 2	C seems to make the most sense as knowing whether buyers actually saw ads and were persuaded by them would help answer the question
Student 1	B won out for me because it could explain why they had more success.
Student 3	It states evaluating the argument, so I think C is correct.
Student 3	The argument is about the advertising firms success
Student 2	I can see what your thinking with B but when I really think about it I think C answers the question
Student 3	Yeah, B and C roughly address the same argument, but C supposrts it more.
Student 1	Seeing if they did actually see the ads would be good... knowing though which ads are more successful might also be good...
Student 2	I was also thinking I don't usually see car ads that contain addresses
Student 3	yeah
Student 1	c
	Below, Student 1 initially does not have the correct choice.
Student 3	I said e, and I'm sticking to it
Student 2	I also think e
Student 2	cause with bolivia they could still have 60 percent like it
Student 2	but with czech, 100 percent could hate it
Student 2	they have no way of knowing
Student 3	And in bolivia they at least had SOME data
Student 3	So they could make an informed choice
Student 2	yeah, totally
Student 1	i picked c but i can see why you guys pick e
Student 2	i think e is the easiest for me to see why it works
Student 2	i don't know if that makes sense
Student 2	it did in my head
Student 3	I have no idea if the czech have stronger money.
Student 1	true, i might change my answer then

Table 1. Two example discussions from E1, assigned the highest rating (4), each with three workers, in bonus condition, where initially two workers were correct and the third switched to the correct answer.

Results

All of the study predictions were supported. Results are reported for 267 worker sessions lasting on average 12.8 minutes (15.0 minutes excluding solo workers), with 169 solo chat sessions, 25 discussions of size 2, and 73 discussions of size 3. About 58% of the 433 attempts to answer questions were answered correctly overall (after discussion), suggesting that the questions were suitably but not too challenging.

138 workers (61%) kept their original choices unchanged on both questions, 74 (33%) changed one answer after the discussion, and 14 (6%) changed both. We found 51% of workers who changed their answers improved their score, while only 18% lowered their score; 86% of workers who changed both answers improved their score. These figures include solo and grouped workers.

We use Fisher's exact test (two-tailed) to examine the connection between independent variables and the proportion of correct answers, because it is appropriate for two-valued outcomes and because it is more conservative than the χ^2 test, particularly for small or unbalanced samples. While we are aware of sophisticated approaches such as Bayesian knowledge tracing [8] that also model whether students might get correct answers by guessing or incorrect answers by slipping, we are interested in the difference between two groups for which such effects should be similar.

Engaging in discussion leads to more correct answers. As Figure 4 (left) shows, among the 169 responses provided by solo crowd workers, 84 (50%) were correct. Among the 269 responses from crowd workers who participated in a discussion group, 169 (63%) were correct, a strongly significant difference (Fisher's exact test, $p < 0.01$).

More discussion is correlated with more correct answers. Each response is associated with a question, a participant, and a particular chatroom discussion. We examined for each response how many chat messages the crowd worker sent in the associated chatroom, and for each possible number of chat messages, determined the proportion of responses that were correct. Sending more chat messages was correlated with a higher proportion of correct answers ($r^2 = 0.53$); this was the case both with the final answers made after the discussion, and the first answers made before the discussion, suggesting correlation rather than causation.

Among workers in dyads and triads, the bonus incentive leads to more correct changed answers. In the no-bonus (control) condition, participants changed 38 out of 127 individual answers (30%); in the bonus condition they changed 47 out of 142 answers (33%). No significant difference was found (Fisher's exact test, $p = 0.60$). However, among the changed answers, 14 answers (11%) were changed from incorrect to correct in the control condition, while 31 (22%) were changed from incorrect to correct in the bonus condition, a significant difference (Fisher's exact test, $p < 0.03$). Among solo workers, changing responses was rare, occurring in only 16 or 9% of 169 solo chat sessions, and changing to the correct response occurred the same number of times (4) in both the bonus and control conditions.

The participants have substantive discussions. Three independent raters applied the following rubric to the discussions:

1. Irrelevant or no discussion: workers did not mention their answers or the question at all.
2. Merely stated answers: One or more workers mentioned their answers, and/or suggested changing answers, but gave no justification or reasoning.
3. Justified answers: One or more students justified their answers, but did not consider or address others' positions.
4. Debate: One or more students justified their answers, and one or more students did at least one of:
   - Considered and addressed positions of other students,
   - Considered and addressed other possible responses, or
   - Persuaded or attempted to persuade other students to change their choice through rational argument.

73 of 98 discussions (74%) were rated 4 by all raters, and 80 (82%) had a median rating of 4. Inter-rater reliability was moderate (Spearman's $\rho = 0.65$). Discussions with 3 participants were generally rated higher than ones with only two, as shown in Figure 4 (right). Table 1 shows two example discussions in which 2 workers begin with the correct answer and the third changes to the correct answer after the discussion.

EXPERIMENT 2

In Experiment 1, workers are asked to discuss the target essay and the five possible answers and try to determine which is the correct answer in a workflow that we call Discuss Question (see Figure 5). This takes place without provision of any kind of instructional material. Workers only choose answers and then discuss their answers with one another.

For Experiment 2, we wanted to see if additional structure would draw out further improvements or changes in the workers' behavior and scores. Our approach was motivated by pedagogy advocated by King [27, 26] which has found that students benefit by generating questions before the group discussion begins, including metacognition monitoring [18]. King's suggested stimuli (questions stems such as "What would happen if ..." and "Compare ... and ... with regard to ...") are not suitable for the GMAT-based tasks, so we designed an alternative set of study materials which we call a *Mini-Lesson*. Workers were shown instructional materials (see Figure 6) and an example critical reasoning problem and asked to identify the unstated assumption that might undermine the argument.

In the study design, we contrasted discussing just this Mini-Lesson with having a second follow-on discussion about the multiple-choice questions directly. Ideally, from a pedagogical perspective, the deeper thinking that comes from examining the assumptions helps obtain the correct answer *before* seeing the choices for the multiple choice question. If group discussion benefits understanding, those who discuss the assumptions should perform better on the multiple choice question than those who generate the assumptions working solo.

Study Design

Sessions: Building on the results of Experiment 1, all groups were of size 1 or 3 (no dyads) and all workers were offered a bonus (solo workers for getting the answer alone, triads if all 3 workers got the right answer). In a between-participants design, workers were assigned to one of three conditions (see Figure 5):

- Minimal: (solo) Read essay, select 1 of 5 answers.
- Mini-Lesson (ML): (triad or solo) Read instructional material, read GMAT essay, generate assumption, discuss assumption, select 1 of 5 answers, write a justification for the answer, view the correct answer.
- Mini-Lesson + Discuss Question (MLDQ): (triad or solo) Read instructional material, read GMAT essay, write down assumption, discuss assumption, select 1 of 5 answers, write justification for the selected answer, discuss justification and answer with group members, select final answer, view correct answer with explanation.

Each participant completed two questions in sequence in their respective condition, with the exception that the mini-lesson was only shown one time. Table 2 shows example discussions for the MLDQ workflow.

Instructions: First, workers were shown the instructional materials and the first problem essay. Then they were asked to identify "an important unstated assumption" in the essay

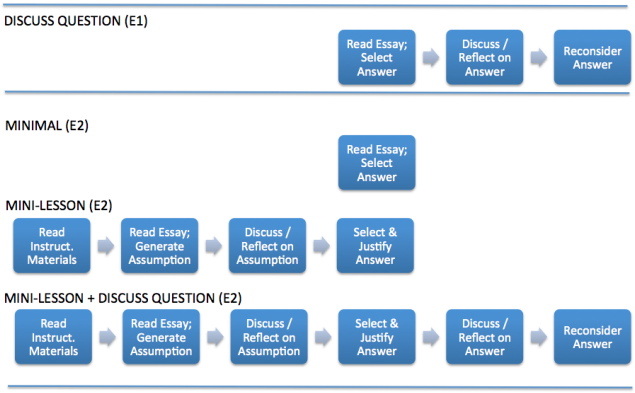

Figure 5. Workflow diagrams for Experiment 1 (E1) and 2 (E2). Not shown are the second questions in the series.

and told the response would be shared with other students on the next screen. Upon entering the discussion, workers were asked to "discuss these assumptions in order to prepare for a question shown on the next screen." Each worker in ML and MLDQ saw a second GMAT question, selected an answer, and provided justifications for responses, but did not have discussions and did not provide assumptions. They were also required to fill out a survey.

Timing: For discussion phases workers were allocated up to 300 seconds; if all members of a chat room signaled that they wanted to end the discussion, then it ended earlier. For choosing and justifying an answer, 240 seconds was allowed, but could end earlier.

Materials: The study materials were a smaller subset of the questions used in E1.

> In this task, you'll learn a critical reasoning skill commonly called identifying hidden assumptions. Read the following carefully. You see arguments every day in statements by politicians, businesses, and advertisements. These arguments are a way of trying to convince you that something is true. Here is an example:
> *Many mobile phone companies protect their technology with patents, and mobile phones continue to become more technically advanced every year. Therefore, patents lead to strong competition in the marketplace.*
> This statement contains the key components of an argument: at least one premise and a conclusion. But often there are gaps in logic between the premises and the conclusion. These gaps can include assumptions, and identifying the unstated assumptions helps you spot the logical flaw in the reasoning: Premise(s) + Assumption(s) = Conclusion(s)
> If the assumptions are incorrect, then the conclusion might not follow. There are several unstated or hidden assumptions in the example above which may be not actually be true:
> **Assumption**: we have plenty of innovation in the presence of patents.
> Perhaps defending against patents takes away resources from making even more innovative inventions.
> **Assumption**: strong competition is the biggest reason for the advanced technology in the mobile phone business.
> However, patents award a monopoly (exclusive rights) to the patent holder, thus reducing competition.

Figure 6. Text of Mini-Lesson given to crowd workers.

Results

This section outlines: descriptive statistics summarizing number of workers in each condition and time spent on the task, the main results regarding which conditions led to an improvement in performance on the two questions, the relationship between survey results and performance, and analyses of various metrics for discussion quality and their relationship to performance.

Condition	Total	Solo	Triad
Minimal	346	346	N/A
Mini-Lesson	409	152	257
Mini-Lesson + Discuss Question	312	122	190
Total	1067	620	447

Table 3. Number of workers in each condition and whether they participated in a discussion group of three (triad).

Condition	Min	Median	Max	IQR	Time limit
Minimal	0.3	2.2	8.8	1.5	9
Mini-Lesson	4.1	13.8	28.4	5.8	30
Mini-Lesson + Discuss Question	8.7	20.3	37.9	6.5	40

Table 4. Overall time spent by workers on the task, in minutes. After the time limit expired, workers could no longer submit.

Descriptive

Worker counts. The three experiment conditions had 346, 409, and 312 workers respectively, for a total of 1067 workers. In conditions with discussion groups, about 60% of workers participated in discussions; see Table 3 for details.

Session times. Worker time increased as stages were added: Minimal flow had a median of 2 minutes, while the Mini-Lesson flow was 14 and Mini-Lesson + Discuss Question was 20 minutes. The maximum time limit for the task was adjusted for each condition accordingly (see Table 4).

Main results

In this experiment, the dependent variables were generally two-valued (whether a response was correct or incorrect), so we rely on two main statistical tools: Fisher's exact test to determine significance, and the 95% confidence interval for the odds ratio, commonly used in medical research, to describe effect size. For example, stating that "group A was 3 to 7 times more likely to answer the question correct than group B" means that there is a 95% chance that the actual odds ratio in the population is between 3 and 7 (i.e. the confidence interval is [3,7]). Because we perform many Fisher tests below, we conservatively selected a significance threshold α using the Bonferroni correction: at least 20 tests were performed, suggesting a rough threshold value of $\alpha = 0.05/20 = 0.0025$.

Mini-Lesson instruction improves performance. We compared solo workers in either ML or MLDQ to the solo workers in the Minimal flow who received no instructional materials. Workers are 7 to 17 times more likely to answer the first question correctly on their first try given the Mini-Lesson scaffolding ($p < 0.0001$). The percentage of correct responses rose from 11% to 58%. Certain questions showed especially dramatic differences, e.g., scores on question 4 rose from 0% of 30 to 82% of 44 workers.

After completing all stages on the first question, workers in scaffolded conditions also go on to perform better independently on the second question. Solo workers in ML were 3 to 7 times more likely to answer the second GMAT question correctly than Minimal flow workers ($p < 0.0001$), with correct responses rising from 20% to 54%. Solo workers in MLDQ, who additionally received an opportunity to reflect on their response to the first question and revise it, were 4 to 9 times more likely to answer the second question correctly than Minimal flow workers ($p < 0.0001$), and correct responses rose from 20% to 60%. Question 4 again exhibits an unusual increase from 5% to 64% for ML, and 5% to 91% for MLDQ.

No significant benefit from generated assumption discussion. The mini-lesson teaches concepts by having workers formulate unstated assumptions about the problem on their own and then discuss these in groups. To justify the discussion component's inclusion, we examine whether it improves performance. We compare ML singleton workers who progressed through the stages alone to ML workers who participated in discussion triads, but found no significant difference in their responses to the first question (59.1% vs. 58.6%, $p > 0.9$ with Fisher's exact test, odds ratio 95% confidence interval $[0.7, 1.5]$) or the second question (54% vs. 56%, $p > 0.7$, $[0.7, 1.6]$). In fact, even when ML singletons are compared to grouped workers in MLDQ, which had the benefit of two discussions, no significant difference could be shown (59% vs. 64%, $p > 0.3$, $[0.8, 2.0]$).

Revising responses

Revised answers after discussion showed improvements. The MLDQ flow provides the opportunity for workers to discuss the answer choices for the first question and then revise them. This is the core feature expected to exploit the benefits of peer learning to improve performance. This process was found to be beneficial: 61% of grouped workers in ML or MLDQ answered the first question correctly on their first try, while 74% of grouped workers in MLDQ answered it correctly following discussion on their initial responses. They were 1.2 to 2.6 times more likely to answer it correctly on their second try ($p < 0.002$).

Among the 190 MLDQ workers who participated in discussions, 45 (24%) changed their response, and of these, 28 or 62.2% revised from an incorrect to a correct response. Solo workers almost never made such positive revisions to their responses: MLDQ workers not participating in groups rarely changed their response (20 of 122 or 16%) and rarely improved their response when doing so (only 4 workers).

Survey feedback

Instructors may be reluctant to deploy new instructional tools that students dislike. To investigate subjective responses to small group discussions, workers were asked to rate their experiences with the discussion. As shown in Figure 7, 401 (53%) of agreed or strongly agreed that the discussion was enjoyable, while only 81 (11%) expressed negative responses.

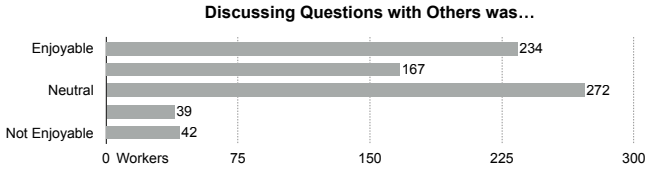

Figure 7. Enjoyment of discussion from Experiment 2 surveys. "Neutral" was chosen most, but more workers enjoyed discussion than did not enjoy it.

In an optional open-ended field for feedback, responses again reflected enjoyment of the task:

- 254 (34%) described the activity as 'interesting', 'fun', or 'enjoyable';
- 12 (1.6%) described the task as 'difficult' or 'hard' *without* also describing it as enjoyable; two of these said it was difficult because they were singletons ("there was no one to chat with");
- 4 described it as 'poor' or 'bad', in all cases because they had no one to chat with or because they thought the other workers were bots (automated responses).

Workers familiar with logical reasoning performed better. Intuitively, workers with a background in the subject area tested by the questions should perform better. The survey asked workers whether they had previously learned about the topic on which the tool provides instruction. Among all ML and MLDQ workers (no instruction was provided in the Minimal flow), 318 (44%) were familiar with the topic and 403 (56%) were not. Workers familiar with the topic were 1.4 to 2.7 times more likely to answer the first question correctly on their first try, and 1.5 to 2.8 times more likely to answer the second question correctly (odds ratio 95% confidence intervals, both $p < 0.0001$ with Fisher's exact test).

No evidence discussion helped workers unfamiliar with logical reasoning. Earlier we found no evidence that group interactions improved performance on the initial response to the first question or on the second question. However, many workers reported prior experience and experienced workers have less room to improve; so we repeat these tests considering only inexperienced workers. There is still no evidence of improvement (61% of ML flow singletons vs. 74% of MLDQ grouped workers, $p > 0.1$ with Fisher's exact test, odds ratio 95% confidence interval [0.9, 3.8]).

Better English speakers perform better. Both logical reasoning and informal discussion can be challenging for workers with limited skills with the language being used, and Turk demographics include many non-native speakers. As expected, workers who rated their English "Very Good" or better were 1.7 to 7.4 times more likely to get the first question correct on their first try than those rating their English lower ($p < 0.001$). There were 719 workers who rated their English "Very Good" or better, and they were correct in 58% of responses, whereas the 35 workers who rated their English lower were correct in only 29%.

Discussion ratings

Because discussion quality varies widely from one discussion to another, and this may influence how helpful it is, we sought to investigate how discussion quality affects performance. We asked two raters to manually label each worker according to their participation in each discussion they participated in. A score between 0 and 3 was the computed for each worker-discussion pair based on the following rubric:

- (1 point) Makes at least one substantive statement;
- (1 point) Makes at least two substantive statements;
- (1 point) Reacts in a substantive manner to a comment made by another at least once.

The subset that was rated included a prefix of the Mini-Lesson (ML) flow discussions, plus a random sample of all other discussions. Workers who were rated by only one rater received that rater's score, while workers rated by both raters received the mean of the two scores. Of the 715 rated workers, 211 were rated 2-3 and 504 were rated 0-1.

No evidence that workers with more substantive participation in discussions perform better. We might expect that workers who are more engaged with the discussion will benefit more from it. However, we found no evidence that workers with 2-3 ratings correctly answered the first question on their first try more often than workers with lower ratings (56% vs. 62%, $p > 0.1$, odds ratio 95% confidence interval $[0.6, 1.1]$). There was also no evidence they performed better on the second question (53% vs. 59%, $p > 0.1$, $[0.6, 1.1]$).

No evidence that high-quality discussions lead to better performance. Even a passive worker can benefit if their group's discussion was fruitful. To analyze this possibility, we summed the ratings of all workers in each discussion, and used this 0-9 score as a metric of overall quality of the discussion. We found no evidence that workers who participated in discussions with scores of 5-9 (29 workers) performed better on the first question on their first try than workers with scores of 0-4 (689 workers) (62% vs. 60%, $p > 0.9$, $[0.5, 2.3]$). There was also no evidence that workers in high-quality discussions performed better on the second question (66% vs. 57%, $p > 0.4$, $[0.7, 3.2]$).

Earlier we compared ML singleton workers and ML grouped workers to determine whether the peer discussion on unstated assumptions was useful. No evidence could be found, and one possible explanation is that not enough discussions were substantive enough to be helpful. Revisiting this, but restricting grouped workers to those who engaged in a discussion rated 5 or higher, we again found no evidence that grouped workers performed better on either the initial response to the first question (55% vs. 59%, $p > 0.8$, $[0.3, 2.2]$) or the second question (60% vs. 54%, $p > 0.6$, $[0.5, 3.3]$).

DISCUSSION

Comparing the Two Experiments

In both Experiments E1 and E2, workers in groups obtained better outcomes in terms of number of correct answers when discussing the actual question prompts than those working

alone. However, in E2, studying the Mini-Lesson and generating a hidden assumption seemed sufficient to produce better outcomes than simply answering the question (as done in the Minimal condition) without the need for group discussion. Those discussions were not rated as highly as the discussions in E1; it may be that the work of generating the assumptions was just as effective as the group discussion in focusing the students on understanding the critical reasoning problem.

Innovating in Group Formation

Research suggests that, for in-person group problem solving, the interaction dynamics of a group can determine the success of group outcomes, independent of the capabilities of the individual group members[61]. An interesting observation from both E1 in the bonus condition and E2 in the MLDQ condition, is that that participants were more likely to move to the correct answer than to an incorrect answer, even if the majority was at the incorrect answer, so long as at least one person was right. The trend was even stronger when the room began with two workers correct.

We also noticed that when all three workers begin with the same answer, the discussion rarely moves them away from that answer even if it is incorrect. This suggests a need for intervention when the discussion group begins on the wrong track. (Other research has found that intervening in poorly performing discussions may work better than in successful ones [53].)

To address this issue, we can take this experimental platform beyond physical classrooms by "virtually" moving participants around to different discussion groups to group them with peers who have stronger understanding of the material. Students who choose an incorrect answer would be grouped with students who have chosen a correct answer to see if learning gains can be achieved. Alternatively, students who all have chosen the same incorrect answer could be given a hint by the system that their answer is incorrect and a suggestion that this guide the conversation. After the students make their next choice, the conversation room they end up in would be determined by which choice they make next in a tree-like structure. Students would continue to end up in new discussion rooms until they have arrived at the correct choice, perhaps with extensive guidance by the system or a suggestion that learning materials be revisited if they get too near the leaves of the tree.

Applicability of Results to MOOCs

Crowd Worker vs. MOOC Learner Demographics

How do the characteristics and motivations of Turkers compare to those of MOOC learners? The two cohorts are similar in many ways:

Environment: like Turkers, MOOC learners are rarely co-located in space or time and know few of their peers.

Education and Demographics: While many educators and commentators justifiably lament the fact that current MOOCs do not serve a wider audience, today's Turkers are intellectually and demographically comparable to MOOC learners. Comparing an earlier study of self-reported Turker demographics [22] and the descriptive statistics of sixteen MIT and Harvard MOOCs in 2012–2013 [19], 55% of US Turkers and about 75% of Indian Turkers had a bachelors' degree or higher, compared with 66% of MOOCers (averaged across all sixteen courses). The median age is 35–39 for US Turkers, 30–34 for Indian Turkers, and 28 for the MOOCers. On three classic critical-thinking problems, Turkers' performance was found to be statistically the same as that of a group of students recruited from a midwestern US university with respect to task completion rate and effect sizes of factors such as outcome bias, conjunction fallacy, and so on [44].

Positive Reception in a Very Large Hybrid Course

In the summer of 2014, the tool was deployed in an introductory undergraduate engineering course of more than 1000 students that is taught in a flipped-classroom style with an online component. Each week the students watched a video and read the corresponding textbook extract, then practiced with multiple choice questions on their own to gauge their level of achievement. They also participated in a two-hour hands-on workshop where they were able to engage in concept checking. A mandatory summative assessment was conducted at the end of the week; using MOOChat on one of two multiple choice questions was the first of two required tasks.

A/B comparisons were not made, but a cursory look at the chat transcripts suggests that they are similar in quality to those of the E1 crowdworker study. An optional questionnaire was administered after the MOOCChat activity containing the prompt "This activity was:" with a 5 point scale where 1 = Not Enjoyable, 3 = Neutral, and 5 = Enjoyable. 615 students responded at least one time (only first answers are reported when repeats occurred). Of these, 53% of students marked Enjoyable (4 or 5), 33% marked Neutral, and 14% marked Not Enjoyable (1 or 2).

This course differs from a MOOC in that the students were collocated in space and time. Although the large size of the class makes it unlikely that any arbitrary triad of students knows one another, it is much easier to ask a group to collaborate simultaneously than in a globally distributed MOOC, and the expectations of participation an in-person course still differ from MOOCs. Therefore, the high participation and positive reception by students in the this course may not be replicated in a standard MOOC.

Evidence of Positive Reception in a MOOC:

We surveyed sixteen learners in a programming-related MOOC [33] after an early "shake-out" deployment of the system to get subjective responses to the tool. 14 of 16 agreed or strongly agreed with the statement "I liked discussing questions in a small group and would like to do so again". Those in groups of size 2 preferred more people in the discussion, whereas those in groups of size 3 indicated that their group size was good. We therefore have reason to expect a positive reception to our approach in a real MOOC deployment.

Counter-Evidence

While we believe this evidence suggests that crowd workers are suitable stand-ins for MOOC learners for the purposes of

testing this approach, there are also important differences for deployment in a real MOOC, including:

Course Context: Unlike crowd workers, MOOC learners would participate in a series of activities covering a set of related topics presented over time in the context of the course's larger narrative, rather than only a single activity. Also, such activities would be just one of several forms of knowledge construction, and the learners would presumably have multiple opportunities to demonstrate retention or transfer of any new learning acquired during the activities.

Motivation to participate: There is a question about how willing students will be to participate in small online group interaction within MOOCs. Although the hundreds of citations in the peer learning literature find that student morale, retention, and learning is increased by this form of person-to-person interaction, we do not know if it will transfer in the less personal, more distributed format of the online chat described here. MOOC students can freely select which course activities they participate in, and their motivations may not match either crowd workers or classroom students. Prior work has found that optional communication channels that expose students' comments to thousands of people such as forums [36, 20] or MOOC-wide chat [6] only attract a fraction of course students. We hypothesize that many students who are not comfortable making posts that are visible to thousands of people will nonetheless be comfortable and even take pleasure in talking in a small group of two or three fellow learners. This would be consistent with advice given to classroom instructors that "Students find it easier to speak to groups of three or four than to an entire class" [11, p. 77].

On the other hand, students in a MOOC do not have to complete all of the work to achieve a passing grade, and so we suspect that adoption will be aided if an instructor makes the use of small-group discussions integral to the running of the course in the same way that peer assessment has been used in some MOOCs [29].

Coordination of group formation: It may be more difficult to coordinate students in MOOCs than in crowd work because of the desire to align every student with a group, as opposed to grouping whichever subset of workers happens to be available at a given time. Instructors will most likely have to encourage students to arrive within pre-stated time periods and it may be necessary to modify MOOC pages to alert students to upcoming group activities.

Implications for Crowd Work

The results here have implications for crowd work. Recall from the related work section that Zhu et al. [62] found that pooling and averaging the results of individual workers improved results on the tasks they studied with the one exception being solving the mathematics problem, where (semi-synchronous) discussion led to better overall results. Thus our results find further evidence that discussion in order to solving problems, as opposed to tasks like brain storming or summarizing, may be a better approach for crowd sourcing.

Zhu et al. also found that reviewing the work of others can improve subsequent task performance. We introduce a different preparatory activity – abstract training and discussion in a mini lesson, which also helped task performance. Future work could directly compare these two approaches.

These results also build on those of Dow et al. [13] which finds that self-assessments in crowd work can lead to better quality results than no feedback; crowd workers in our study who were asked to reflect on their choices in a chat room alone achieved improvements over those who did not do so.

More generally, these results suggest that looking more deeply into the learning literature may yield additional ways to improve crowd sourcing approaches and outcomes.

CONCLUSIONS AND FUTURE WORK

We conducted two sets of experiments to bring peer learning into synchronous group work at scale. The first experiment tested unstructured discussion, compared different size groups, and tested the effects of a bonus condition. This study found evidence that discussing challenging questions in an online chat can lead to participants converging to the correct answer more often than answering the question alone. The study also found that participants were more motivated to get the correct results when an extrinsic reward was offered. These results mirror what has been found in in-person classrooms and suggests that by properly incentivizing students to participate in discussions, peer learning can be successfully introduced into MOOCs. Forming synchronous discussion groups at fixed time intervals is one way to overcome the inherent asynchrony in MOOCs though there may be alternative effective group formation methods.

The second set of experiments tested the effects of adding more structure to the interaction among students, again following the pedagogy of peer instruction, finding that adding progressively more structure and more opportunities for discussion to the learning experience led to better outcomes.

Although we do not claim that participants in this study were learning in the same way that happens in a classroom, we do think these results provides evidence that such an approach may be helpful in a MOOC to improve learning outcomes. Many crowd workers commented about enjoying the task, either in the chatroom or in the freeform section of the questionnaire. A comment made during one discussion was "wish they had these kinds of things when I was in school", and an initial very small pilot with MOOCs positive subjective reactions to the approach. This also reflects classroom findings that many students prefer the social nature of interacting with others while learning, and may bode well for improving retention in MOOCs.

Future work will include deployment in MOOCs and we are actively seeking collaborating instructors. In the coming weeks the system will be used for practice questions for an exam in an online section of an in-person course consisting of more than 1,000 students. The results of this exercise will shed additional light on the applicability of this approach to real courses.

Students individually write down unstated assumptions:	
Student 1	File sharing might be done between different people as well as different devices.
Student 2	This assumes that everyone sharing files must also be purchasing the CDs.
Student 3	That "illegal" file sharing actually helps cd.
Discussion of assumptions begins:	
Student 2	I mean, I think we're all pretty right.
Student 3	I think so too
Student 2	cuz in order to help CD sales, those illegally sharing them would have to be purchasing them to share
Student 2	but I feel like that's not what happens. I feel like one person would share and then that one would keep getting shared
Student 3	And if someone liked what they were hearing they would be open to purchasing the artist
Student 3	if they "careD"
Student 2	right, exactly
Student 2	I feel like in one way it helps broaden appreciation cuz you can get so much of it
Student 1	yeah
Student 2	then again, people at my university would just interlibrary loan CDs and then copy them to their computers.
Student 2	"the cd was bought at one point!"
Student 1	well, I would agree with that
Student 1	but I think most university students are like that
Student 3	i think that happens at every college campus
Student 2	I'm sure it does hahahaa
Student 2	that and stealing from the dining halls
Student 1	*gasp* no
Student 3	Never!
Student 2	"we already pay so much for tuition!"
Student 2	oh, entitlement
Discussion ends; students are shown 5 choices, they individually make a choice and write a justification for their answer:	
Student 1	They don't know for sure that an increase in music CD sales is due to file sharing. It could be due to any number of things (better music choices, different groups of people buying CDs than in the past, easier to purchase CDs now than before - online vs. physical store, etc.).
Student 2	Because it assumes that two things are linked just because they have to do with music, while the people sharing finals and the people who are buying CDs could be completely different people.
Student 3	Some people will never pay what they can get for free
Students 1&2 agree on the answer; Student 3 differs. Discussion of answers begins:	
Student 1	BAM
Student 2	So do we all pick D? and get $1?
Student 3	Thats essentially what it looks like
Student 1	well, we all have to get the same answer
Student 1	to get the bous
Student 2	And it has to be correct
Student 1	assuming that choice D is "correct"
Student 3	I was stuck between B, and D
Student 3	so I can swing both ways, literally
Student 2	I think it's D because it thinks that just relating to music is enough of a connection
Student 2	when the people who buy CDs may not be the ones downloading/sharing
Student 1	same here
Student 1	I think it's D
Student 2	So do we all want to stick with D
Student 3	I shall switch over to D
Student 2	cool
Student 1	aight

Table 2. Example assumptions, justifications and discussion from E2 MLDQ condition.

ACKNOWLEDGEMENTS

We thank Carl Reidsema, Grant Edwards, and Philip Long of the University of Queensland for partnering with us in the hybrid course experiment, Robert Hernandex and Steven Walentino for their work on labeling the chat sessions, and Magoosh Inc for supplying the GMAT practice questions. This material is based upon work supported by a Google Social Interactions Research Award and the National Science Foundation under Grant No. IIS 1149799 and IIS 1210836.

REFERENCES

1. Astin, A. W. *What matters in college? Four critical years revisited.* Jossey-Bass, 1993.
2. Bernstein, M. S., Brandt, J., Miller, R. C., and Karger, D. R. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of UIST*, ACM (New York, NY, USA, 2011), 33–42.
3. Bonwell, C. C., and Eison, J. A. *Active learning: Creating excitement in the classroom*. School of Education and Human Development, George Washington University Washington, DC, 1991.
4. Chase, J., and Okie, E. G. Combining cooperative learning and peer instruction in introductory computer science. *ACM SIGCSE Bulletin 32*, 1 (2000), 372–376.
5. Chen, W., and Looi, C.-K. Active classroom participation in a group scribbles primary science classroom. *British Journal of Educational Technology 42*, 4 (2011), 676–686.
6. Coetzee, D., Fox, A., Hearst, M. A., and Hartmann, B. Should your MOOC Forum Use a Reputation System? In *CSCW'2014* (2014).
7. Cooper, S., and Sahami, M. Reflections on Stanford's MOOCs. *Commun. ACM 56*, 2 (Feb. 2013), 28–30.
8. Corbett, A. T., and Anderson, J. R. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction 4*, 4 (1994), 253–278.
9. Cottell, P. Cooperative Learning in Accounting, Chapter 2. In *Cooperative Learning in Higher Education: Across the Disciplines, Across the Academy*. Stylus Publishing, LLC., 2012.
10. Crouch, C. H., and Mazur, E. Peer instruction: Ten years of experience and results. *American Journal of Physics 69*, 9 (2001), 970–977.
11. Davis, B. G. *Tools for Teaching*. Jossey-Bass, 1993.
12. Deslauriers, L., Schelew, E., and Wieman, C. Improved learning in a large-enrollment physics class. *Science 332*, 6031 (2011), 862–864.
13. Dow, S., Kulkarni, A., Klemmer, S., and Hartmann, B. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ACM (2012), 1013–1022.
14. Fischer, F., Bruhn, J., Grsel, C., and Mandl, H. Fostering collaborative knowledge construction with visualization tools. *Learning and Instruction 12*, 2 (2002), 213 – 232.
15. Glass, A. L., and Sinha, N. Multiple-choice questioning is an efficient instructional methodology that may be widely implemented in academic courses to improve

exam performance. *Current Directions in Psychological Science 22* (2013), 471.

16. Gweon, G., Rosé, C. P., Carey, R., and Zaiss, Z. S. Providing support for adaptive scripting in an on-line collaborative learning environment. In *CHI 2006* (2006).

17. Hake, R. R. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics 66* (1998), 64.

18. Haller, E. P., Child, D. A., and Walberg, H. J. Can comprehension be taught? a quantitative synthesis of metacognitive studies. *Educational researcher 17*, 9 (1988), 5–8.

19. Ho, A. D., Reich, J., Nesterko, S. O., Seaton, D. T., Mullaney, T., Waldo, J., and Chuang, I. HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013. Tech. rep., Massachusetts Institute of Technology and Harvard University, Cambridge, MA, Jan 2014.

20. Huang, J., Dasgupta, A., Ghosh, A., Manning, J., and Sanders, M. Superposter behavior in mooc forums. In *Proceedings of the first ACM conference on Learning@ scale conference*, ACM (2014), 117–126.

21. III, H. L. R., and Karpicke, J. D. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science 17*, 3 (2006).

22. Ipeirotis, P. G. Demographics of mechanical turk. Tech. rep., New York University Stern School of Business, mar 2010.

23. Isenberg, D. J. Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology 50*, 6 (1986), 1141.

24. James, M. C., and Willoughby, S. Listening to student conversations during clicker questions: What you have not heard might suprise you! *American Journal of Physics 79*, 1 (Jan 2011), 123–132.

25. Johnson, D. W., Johnson, R. T., and Smith, K. A. *Active learning: Cooperation in the college classroom*. Interaction Book Company Edina, MN, 1991.

26. King, A. Improving lecture comprehension: Effects of a metacognitive strategy. *Applied Cognitive Psychology 5*, 4 (1991), 331–346.

27. King, A. Structuring peer interaction to promote high-level cognitive processing. *Theory into practice 41*, 1 (2002), 33–39.

28. Kittur, A., Chi, E. H., and Suh, B. Crowdsourcing user studies with mechanical turk. In *CHI'08* (2008), 453–456.

29. Kulkarni, C., Wei, K. P., Le, U., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., and Klemmer, S. Peer and self assessment in massive online classes. *ACM Transactions of Computer-Human Interaction (TOCHI) 20*, 6 (2013).

30. Lasecki, W. S., Song, Y. C., Kautz, H., and Bigham, J. P. Real-time crowd labeling for deployable activity recognition. In *Proceedings of CSCW*, ACM (New York, NY, USA, 2013), 1203–1212.

31. Laughlin, P. R., and McGlynn, R. P. Collective induction: Mutual group and individual influence by exchange of hypotheses and evidence. *Journal of Experimental Social Psychology 22*, 6 (1986), 567–589.

32. Lieberman, M. D. Learning from others. *The Chronicle of Higher Education* (April 2014), B4–B5.

33. Lim, S., Coetzee, D., Hartmann, B., Fox, A., and Hearst, M. A. Initial experiences with small group discussions in moocs. In *Proceedings of the first ACM conference on Learning@ scale conference*, ACM (2014), 151–152.

34. Lord, T. R. 101 reasons for using cooperative learning in biology teaching. *The American Biology Teacher 63*, 1 (2001), 30–38.

35. Magee, R. M., Mascaro, C. M., and Stahl, G. Designing for group math discourse. In *2013 Conference on Computer Supported Collaborative Learning (CSCL 2013)* (2013).

36. Mak, S., Williams, R., and Mackness, J. Blogs and forums as communication and learning tools in a MOOC. In *International Conference on Networked Learning 2010* (2010), 275–285.

37. Martin, F. G. Will massive open online courses change how we teach? *Commun. ACM 55*, 8 (Aug. 2012), 26–28.

38. Martin, F. G. Will MOOCs change how we teach? *CACM 55*, 8 (aug 2012).

39. Mazur, E. *Peer Instruction: A User's Manual*. Prentice-Hall, Upper Saddle River, NJ, 1991.

40. Millis, B. *Cooperative Learning in Higher Education: Across the Disciplines, Across the Academy*. Stylus Publishing, LLC., 2012.

41. Millis, B. J., and Cottell, P. G. *Cooperative learning for higher education faculty*. Oryx Press (Phoenix, Ariz.), 1998.

42. Mullen, B., Johnson, C., and Salas, E. Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and applied social psychology 12*, 1 (1991), 3–23.

43. Palinscar, A. S., and Herrenkohl, L. R. Designing collaborative contexts: Lessons from three research programs. In *Cognitive perspectives on peer learning*. Lawrence Erlbaum Associates Publishers, 1999.

44. Paolacci, G., Chandler, J., and Ipeirotis, P. G. Running experiments on amazon mechanical turk. *Judgment and Decision making 5*, 5 (2010), 411–419.

45. Robinson, P., and Cooper, J. L. The Interactive Lecture in a Research Methods and Statistics Class, Chapter 7. In *Cooperative Learning in Higher Education: Across the Disciplines, Across the Academy*. Stylus Publishing, LLC., 2012.

46. Ruhl, K. L., Hughes, C. A., and Schloss, P. J. Using the pause procedure to enhance lecture recall. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children 10*, 1 (1987), 14–18.

47. Sadler, P. M., and Good, E. The impact of self-and peer-grading on student learning. *Educational assessment 11*, 1 (2006), 1–31.

48. Setlock, L. D., Fussell, S. R., and Neuwirth, C. Taking it out of context: collaborating within and across cultures in face-to-face settings and via instant messaging. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, ACM (2004), 604–613.

49. Smith, K. A., Matusovich, H., Meyers, K., and Mann, L. Preparing the Next Generation of Engineering Educators and Researchers, Chapter 6. In *Cooperative Learning in Higher Education: Across the Disciplines, Across the Academy*. Stylus Publishing, LLC., 2012.

50. Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., and Su, T. T. Why peer discussion improves student performance on in-class concept questions. *Science 323*, 5910 (2009), 122–124.

51. Springer, L., Stanne, M. E., and Donovan, S. S. Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research 69*, 1 (1999), 21–51.

52. Strijbos, J.-W., and Stahl, G. *Analyzing Interactions in CSCL: Methods, Approaches and Issues (Computer-Supported Collaborative Learning Series)*. Springer, 2010, ch. How to Study Group Cognition.

53. Tausczik, Y. R., and Pennebaker, J. W. Improving teamwork using real-time language feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2013), 459–468.

54. Trausan-Matu, S., Dascalu, M., and Rebedea, T. PolyCAFe: Automatic support for the polyphonic analysis of CSCL chats. *International Journal of Computer Supported Collaborative Work (ijCSCL) 9*, 2 (2014).

55. Trytten, D. A. Progressing from small group work to cooperative learning: A case study from computer science*. *Journal of Engineering Education 90*, 1 (2001), 85–91.

56. Walther, J. B. Interpersonal effects in computer-mediated interaction a relational perspective. *Communication Research 19*, 1 (1992), 52–90.

57. Walther, J. B., Loh, T., and Granka, L. Let me count the ways: The interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity. *Journal of Language and Social Psychology 24*, 1 (2005), 36–65.

58. Wang, H.-C., Fussell, S. F., and Setlock, L. D. Cultural difference and adaptation of communication styles in computer-mediated group brainstorming. In *CHI'09* (2009), 669–678.

59. Webb, N. M., and Farivar, S. Developing productive group interaction in middle school mathematics. In *Cognitive perspectives on peer learning*. Lawrence Erlbaum Associates Publishers, 1999.

60. Wilkowski, J., Russell, D. M., and Deutsch, A. Self-evaluation in advanced power searching and mapping with google moocs. *ACM Learning At Scale, L@ S 2014* (2014), 04–05.

61. Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., and Malone, T. W. Evidence for a collective intelligence factor in the performance of human groups. *science 330*, 6004 (2010), 686–688.

62. Zhu, H., Dow, S. P., Kraut, R. E., and Kittur, A. Reviewing versus doing: Learning and performance in crowd assessment. In *CSCW'2014*, ACM (2014).