STRUCTURING PEER INTERACTIONS FOR MASSIVE SCALE LEARNING

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Chinmay Eishan Kulkarni

Aug 2015

Abstract

Massive open online classes (MOOCs) offer an opportunity to dramatically broaden access to education. However, dramatically broadened access also creates challenges. Classes enroll tens of thousands of students, all of whom participate remotely and asynchronously based on their own schedule. This large, asynchronous and remote access in turn makes it challenging to scale effective teaching techniques that rely on personal interactions between teacher and student, such as open-ended assessment and discussion, and rapid formative feedback.

This dissertation brings the benefits of effective teaching techniques to massive online classes, by introducing computational systems that replace hard-to-scale teacher-student interactions with peer interactions. Because peer interactions rely on interactions between students, they can potentially scale to any classroom size. In this dissertation first, I first study the causal mechanisms that lead to the learning benefits of classroom techniques like feedback and discussion. Then, I introduce interfaces that combine these operative mechanisms with the properties of online classes, such as mediated communication and the large number of students.

This dissertation develops these ideas through two large-scale systems, PeerStudio and Talkabout, which target fast, revision-oriented feedback, and global-scale student discussions, respectively. This dissertation also includes the first large-scale evaluation of a global peer-assessment system.

PeerStudio uses the temporal overlap in student schedules at large scale so that students receive fast, revision oriented feedback from classmates at any time of day. Talkabout leverages the globally distributed student participation to create discussions where students speak with peers with diverse experience and viewpoints. Controlled experiments show both systems improve both students' learning experience and their grades.

These systems, and the large-scale evaluations that led to their design, point to a future in which classrooms rely on the collective experiences of their students, and students around the world have access to education that is as effective as it is accessible.

Acknowledgments

If you are reading this, then thanks to the valiant efforts of Gabor Angeli, Jillian Hess Lentz, and the Stanford Registrar’s office, my dissertation was successfully submitted while I was attending the New Faculty Retreat at Carnegie Mellon University. (Gabor and Jillian, Jean Yang's thesis starts similarly; I saw hers and realized this thesis would be literally impossible without you. Thank you!) Now on to other acknowledgments.

When I entered graduate school, Stanford’s Computer Science department had just started its rotation program for new PhD students. This is a fantastic program, and through it, I met not only my advisor Scott Klemmer, but also Hector Molina-Garcia and Jeff Heer. Jeff showed me how the best research could impact the world directly. Hector showed boundless enthusiasm for my research, and I always left his office believing everything was achievable again. When I grow up, I want to be Hector.

Most importantly, Stanford helped me meet Scott Klemmer, who has profoundly shaped how I think about the world. Few mentors are as insightful, far thinking, and organized as Scott is. And few students are as fortunate as I am to have had his advice in all my endeavors. Scott once wrote to me, “I view advising as a lifelong connection—even after you graduate, I'm there when you need it.” I have no doubt I will rely on Scott's advice for years to come.

When Scott moved to San Diego, I was very worried. However, this turned out to be a blessing in disguise. Not only did Scott continue to advise me, I had the opportunity to work with Michael Bernstein. Michael has always been "AWESOME!" (Fellow advisees will recognize this as Michael's favorite line—I have rarely seen anyone as positive!) Michael, thank you for your advice, your time, and most importantly your empathy. It made all the difference!

I would also like to thank Ed Cutrell and Ed Chi, who have been my mentors even before graduate school, and Danyel Fisher, who somehow persuaded Scott that I was worth taking a chance on.

Some people have horror stories about their thesis committee members; I have trouble relating to them. Dan McFarland and Dan Schwartz have been my strongest supporters. I'm especially grateful to Dan McFarland, who read my dissertation while visiting his family so that I could submit it on time.

Stanford has supported me in ways that make colleagues envious. In this regard, I especially want to thank Verna Wong, Peche Turner, and Jillian Lentz for their kind “administrative” support (a better term would be “life-saving.”).

I have had wonderful collaborators—in particular I want to thank Daphne Koller and the warm and wonderful team she has brought together at Coursera. I also want to thank Julia Cambre, Courtney Noh, and Yasmine Kotturi. As undergraduate researchers, I expected to work with them for only a few months, but they have been generous with their time, and I have continued to collaborate with them for more than two years now.

I feel incredibly lucky for how fun and memorable my friends made my life as a graduate student: they are the smartest, most interesting, and most supportive people you can imagine. And finally, I would like to thank my parents, whose constant encouragement, reminders, support, and everyday sacrifices have made this dissertation possible.

Table of Contents

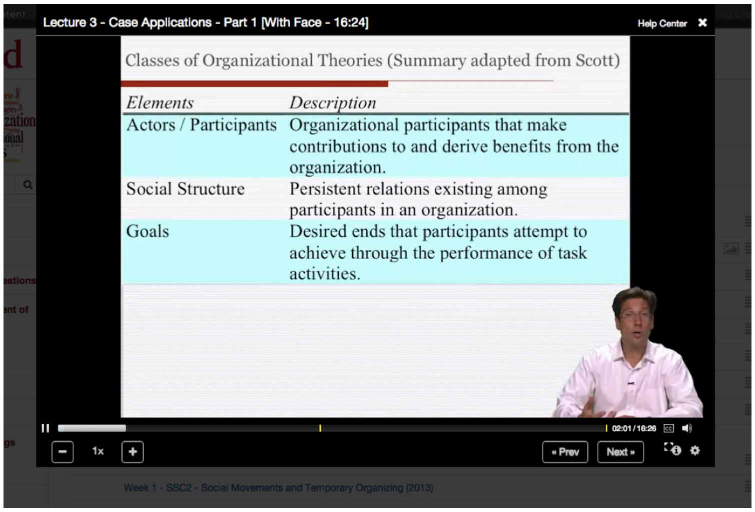
List of Tables

List of Illustrations

Chapter 1
Massive-scale education: Challenges and opportunities

Massive online classes offer an opportunity to dramatically expand access to education. Today, millions of students from all over the world participate in these classes to learn a vast array of topics and skills, such as machine learning, interaction design, modern music, and international relations. Students in these classes watch pre-recorded lectures (Figure 1), discuss questions from class in online forums (Figure 2), and work on assignments assessed by automated systems, all of which are accessible through a web browser.

However, the same broad access that makes MOOCs powerful also creates challenges for traditional teaching methods. The most effective teaching in the physical classroom has relied on personal interactions of students with the teacher and with each other, seeking and providing feedback on open-ended work [1][2], and discussing material [3]. However, in a large online class, tens of thousands of students enroll at once, and they participate in the class remotely and asynchronously based on their own


Lecture 3 - Case Applications - Part 1 [With Face - 16:24]
Help Center
02:01 / 16:26
1x
« Prev
Next »
Week 1 - SSC2 - Social Movements and Temporary Organizing (2013)

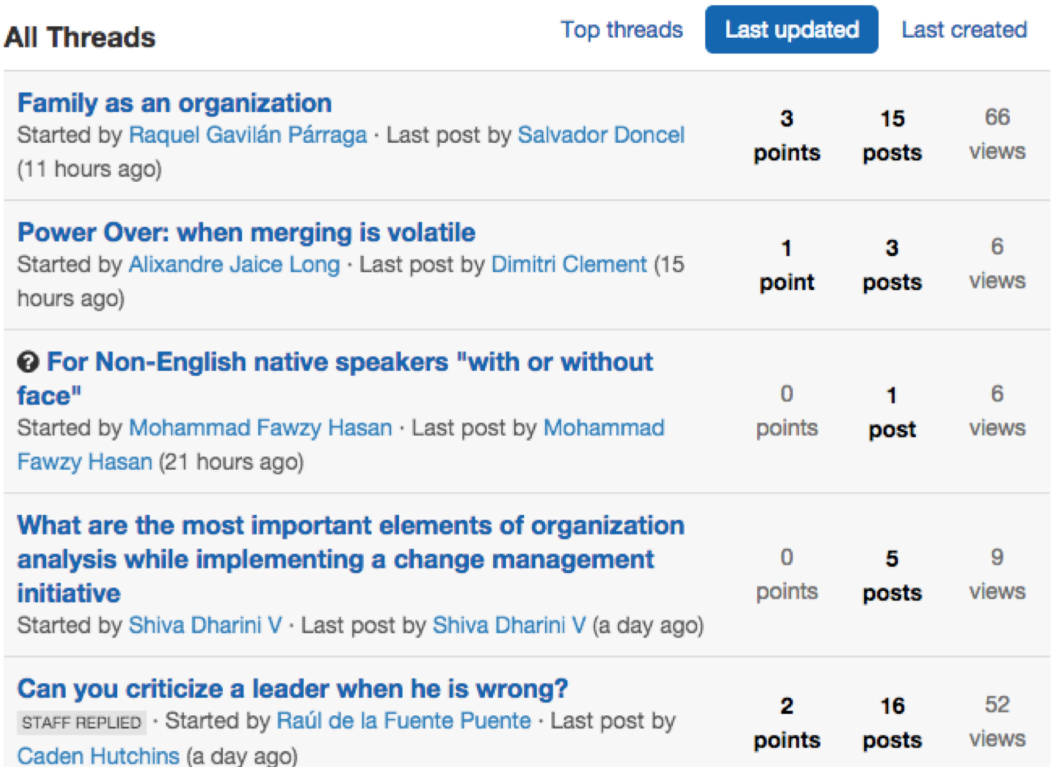
Classes of Organizational Theories (Summary adapted from Scott)

Elements	*Description*
Actors / Participants	Organizational participants that make contributions to and derive benefits from the organization.
Social Structure	Persistent relations existing among participants in an organization.
Goals	Desired ends that participants attempt to achieve through the performance of task activities.

Figure 1: a typical video lecture from an online class, including slides and a professors' explanation.

schedule. This large, asynchronous and remote access makes it challenging to scale effective teaching techniques that rely on personal interactions between teachers and students.

Peer interactions offer a promising alternative to teacher-student interactions. Because peer interactions rely on interactions between students, they are not bottlenecked by the number of teachers, and can potentially scale to any classroom size. Furthermore, interactions between peers improve student motivation; both when interactions are informal and unstructured [4], and when they are formal and structured, such as debate or collaboration [5]. They also lead to greater student inquiry [6] and deeper learning (see Chapter 6). Similarly, open-ended work is crucial for learning in creative domains where students ask their own questions, create their own solutions, and receive feedback from peers and instructors [7].



Family as an organization Started by Raquel Gavilán Párraga · Last post by Salvador Doncel (11 hours ago)	3 points	15 posts	66 views
Power Over: when merging is volatile Started by Alixandre Jaice Long · Last post by Dimitri Clement (15 hours ago)	1 point	3 posts	6 views
? **For Non-English native speakers "with or without face"** Started by Mohammad Fawzy Hasan · Last post by Mohammad Fawzy Hasan (21 hours ago)	0 points	1 post	6 views
What are the most important elements of organization analysis while implementing a change management initiative Started by Shiva Dharini V · Last post by Shiva Dharini V (a day ago)	0 points	5 posts	9 views
Can you criticize a leader when he is wrong? STAFF REPLIED · Started by Raúl de la Fuente Puente · Last post by Caden Hutchins (a day ago)	2 points	16 posts	52 views

Figure 2: Online classes also often include threaded forums for student questions and discussion.

However, peer interactions, as they occur in the physical classroom, are also challenging to scale. These peer interactions usually occur at small scales, synchronously, and with the active involvement of the teacher. How could we scale these effective peer-learning methods from the small classroom to the broad reach and access of an online class?

To effectively scale peer learning to an online classroom, we must solve two challenges. First, students in online classes participate on their own schedule; therefore students should be able to participate asynchronously, while still interacting with peers. Asynchronous access makes broad access possible, but it also makes classroom techniques such as discussion challenging: students in online classes often learn in isolation with class forums representing students' only portal into their global peer group (See Chapter 7). Second, because students participate in online classes remotely, peer interactions must be scaffolded remotely as well. For example, assessing open-ended work requires expertise and training [7]–[9], that has traditionally relied on interacting with teachers directly. In an online class, interfaces need to provide such scaffolding, such as assessing open-ended work and suggesting ways to improve it (Chapter 3).

Motivated by the broad reach and promise of massive online education, this dissertation asks: how might computational systems help solve the two challenges of supporting asynchronous participation and remote scaffolding, to enable large numbers of students to participate in pedagogically useful peer interactions? Furthermore, it asks, could we leverage particular properties of this new environment to create entirely new learning opportunities?

I focus on two particular properties of online classes that may enable us to build large-scale peer interaction systems: computer-mediated peer interactions and global scale. In an online classroom, any collection of peers can interact with each other in real-time regardless of geographical separation. These peer interactions can allows students to learn from peers not in their immediate vicinity Furthermore, interactions can be monitored and structured more precisely than in the physical classroom. With global scale, large numbers of students from around the world participate in the same class at nearly the same time [10]. This large number of globally distributed students allows computational systems that support asynchronous access. Furthermore, global scale leads to a geographically diverse classroom; allowing systems to use global diversity as a pedagogical asset.

Thus, this thesis leverages computer-mediated communications and global scale to create new interactive systems and pedagogical methods. To do so, however, it only scales the mechanisms that lead to peer interactions being effective in the physical classroom, not their specific form. For example, it creates systems for rapid feedback on open-ended work, similar to a design studio, but students receive feedback on a website bearing little resemblance to a design studio (Chapter 5). Preserving the operative mechanisms but redesigning interaction systems and methods allows systems to use the properties of scale as an asset, not just a hindrance.

To understand the mechanisms of peer learning, this dissertation relies heavily on research about how people learn through peer interactions in the small classroom environment (e.g. from prior work in learning science and psychology). This literature informs a recurring theme in this dissertation: the notion of developing pedagogy and software in synchrony. In particular, this thesis develops pedagogy that leverages the massive scale of the online class in structuring peer interactions, and software that thousands of students can reliably use to participate in these structured interactions. For example, this thesis introduces pedagogical techniques that leverage the enormous geographic diversity in an online class through peer discussion. It also introduces Talkabout (software) that chooses maximally diverse participants for such discussion groups, and enables students to interact via video chat.

More than 100,000 students in more than a hundred online classes have used the software systems introduced in this thesis. This corresponds roughly to 50 person-years of non-stop utilization. This wide usage has also allowed me in this thesis to conduct large-scale, online, controlled experiments that extend our knowledge of learning both at large global scales and small local scales.

Below, I describe the main contributions of this thesis organized by the learning interactions they target.

Scaling open-ended assessment

Open-ended work is key to learning in creative domains such as engineering, design, and the arts [7]. Open-ended work requires students to construct a solution, and often to also construct their own problems. For example, students might design a smartphone app in a Human-Computer Interaction class (if students must also decide what application to build, they construct both the problem and the solution); or they might create a threading system in an Operating Systems class. Constructing solutions, rather than simply recognizing correct answers (e.g. with multiple choice questions), enables students to practice complex skills and creativity [7].

Assessing open-ended assignments at large scale is challenging. Open-ended work has many good solutions, and evaluating quality often requires interpretation and judgment. Returning of our examples, teachers might assess if the smartphone app is well designed, or if the threading code is modularized well. This assessment requires both common-sense knowledge to understand student work and the expertise to assess tacit criteria such as “well-designed” or “well-modularized” that cannot be completely articulated. Indeed, teaching such tacit criteria is an important goal in open-ended domains like design [11].

Online classes have thousands of students submitting solutions, but because acquiring broad expertise in most fields takes many years [12], the staff resources for assessing these submissions are limited. Consequently, large classes (especially online, but in-person too) limit themselves to easily verifiable assessments like multiple-choice questions.

This dissertation introduces methods by which *peers* can assess and give feedback on open-ended work at a large scale. One thing that scales naturally with the number of students is the number of classmates who could act as peer raters. But how can peers acquire the necessary expertise in judgment? Informed by systems like FoldIt [13], this thesis leverages the insight that peer assessment could scale by creating *micro-expertise*: students might lack the broad expertise of teaching staff, but equipped with structured rubrics, well-placed examples, and just-in-time assessment training, students could expertly critique a particular assignment.

In this thesis, I describe how I and my collaborators at Coursera created the first MOOC-scale peer assessment platform. Our peer assessment system has enabled assessment of open-ended work in more than a hundred massive classes on Coursera. Students have used it to critique work in disciplines as varied as programming, design, poetry, and finance (Figure 3). In all, more than 100,000 students have used this as-

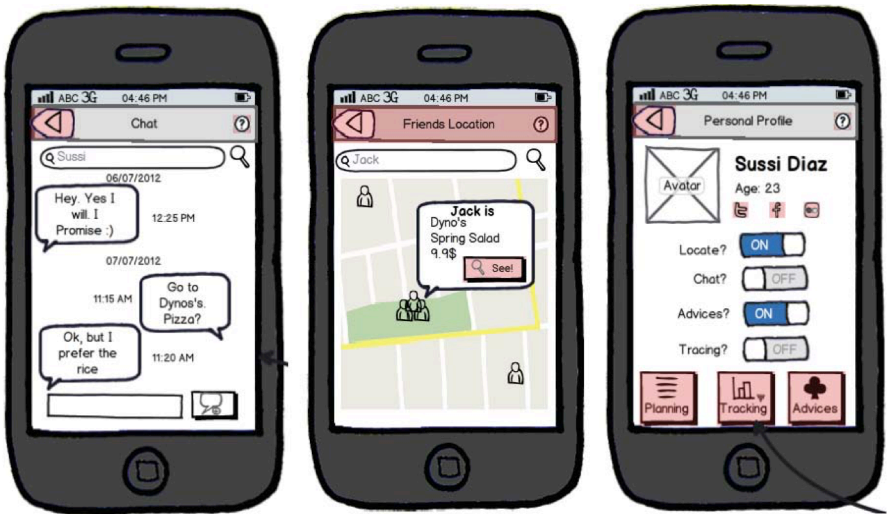


Figure 3: An example student project assessed by peers. Here, peers evaluate the design of a proposed mobile phone application.

sessment system.

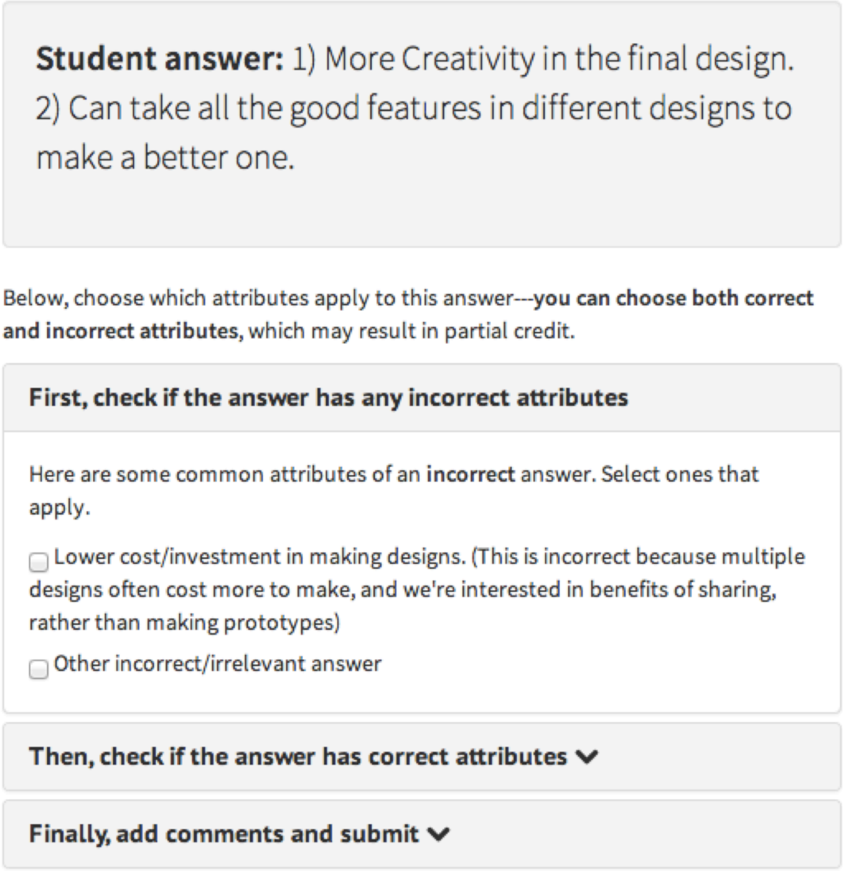
In this thesis, I also contribute the first evaluation of massive-scale peer feedback. A study with approximately 10,000 students found that this system yields accurate grading—97.5% of students who would earn certificates with staff assessment did so, while only 1.2% of students were awarded certificates in error. Furthermore, students reported that peer assessment provided diverse feedback, inspiration, and reflective opportunities.

Since our system was released, many others have built on the concept of online peer assessment. My work assumes that students all assess similarly; by explicitly modeling student biases and uncertainties, others have trained machine-learning models that improve grade agreement with staff [14]. Researchers have also applied the techniques introduced in this thesis for systems that enable designers to seek feedback on prototypes on crowd-sourcing markets like Mechanical Turk [15]. Our datasets of peer assessment have also been used to evaluate algorithms for cost minimization of crowd-sourced rating [16], and for quality control of crowd work [17].

How can machines help minimize grading effort?

Data from our peer assessment system indicated it was accurate, but that assessment required a large amount of student effort. For instance, assessing a single assignment in the human-computer interaction class required 15,000 hours of effort from 4,000 students. Could we structure online interactions so student effort is directed where it matters most?

This chapter demonstrates how to direct students to the peer work that would benefit most from their feedback. To find student work that would benefit most from their feedback, we observe that some student submissions exhibit obvious errors or poor quality, and might need fewer raters to provide useful feedback. Conversely, more raters might be necessary for a novel response. To allocate peer effort, we combined peer assessment with logistic text-classifier models. These models were trained on TA-graded student answers using grades as labels. Since these regression models used simple keyword features, they are accurate for the most common student responses, but not for responses that are novel, and use few keywords present in training data. Thus, the distance of a student's response from the decision boundary (which is dependent on the keywords present) yields an approximate measure of grade uncertainty. We found that compared to spreading reviewer effort evenly, recruiting more peers for



Student answer: 1) More Creativity in the final design. 2) Can take all the good features in different designs to make a better one.

Below, choose which attributes apply to this answer---**you can choose both correct and incorrect attributes,** which may result in partial credit.

First, check if the answer has any incorrect attributes

Here are some common attributes of an **incorrect** answer. Select ones that apply.

□ Lower cost/investment in making designs. (This is incorrect because multiple designs often cost more to make, and we're interested in benefits of sharing, rather than making prototypes)

□ Other incorrect/irrelevant answer

Then, check if the answer has correct attributes ⌄

Finally, add comments and submit ⌄

Figure 4: When students assessed short-answer submissions, they labeled aspects of the answer that were correct and incorrect. This information can be used to provide students early, automated feedback.

answers with uncertain grades and fewer otherwise cut the student effort in half, and still yields grades that are 80-90% as accuracy.

Second, we also used text-classification models to provide students early, machine-generated feedback. When students assessed short-answer submissions, they labeled aspects of the answer that were correct and incorrect (Figure 4). We then built a classifier system trained on these labels that could provide students early feedback.

Overall, our experience suggests that artificial intelligence systems can be beneficially combined with peer interactions, in particular to yield low-accuracy feedback quickly, and to direct peer efforts so high-accuracy feedback is less expensive.

How can peers help students revise and attain mastery?

To gain mastery, students must practice, receive feedback, and revise work [18]. However, in practice, revision is rare both in online classes and in universities. Students need reliably rapid feedback to plan for revisions and have adequate time to revise, which are both typically unavailable. For instance, it can easily take a week to

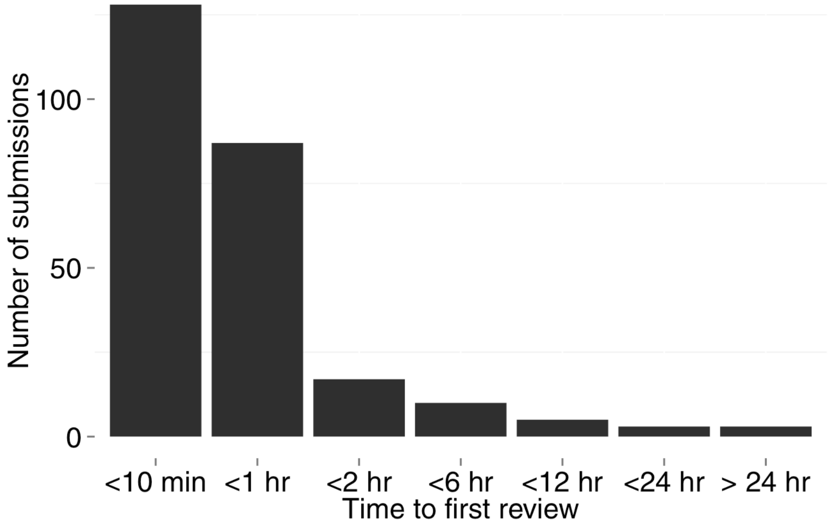


Figure 5: PeerStudio provides rapid peer feedback in MOOCs. The median student receives feedback in 20 minutes after submission

receive instructor or peer feedback on work. Feedback is also often coupled with a summative grade, and classes move to new topics faster than feedback arrives. The result is that many opportunities to develop mastery are lost, as students have little opportunity and no incentive to revise work. Could the scale of an online class enable new opportunities for revision? Furthermore, could students learn from the experience of offering each other feedback?

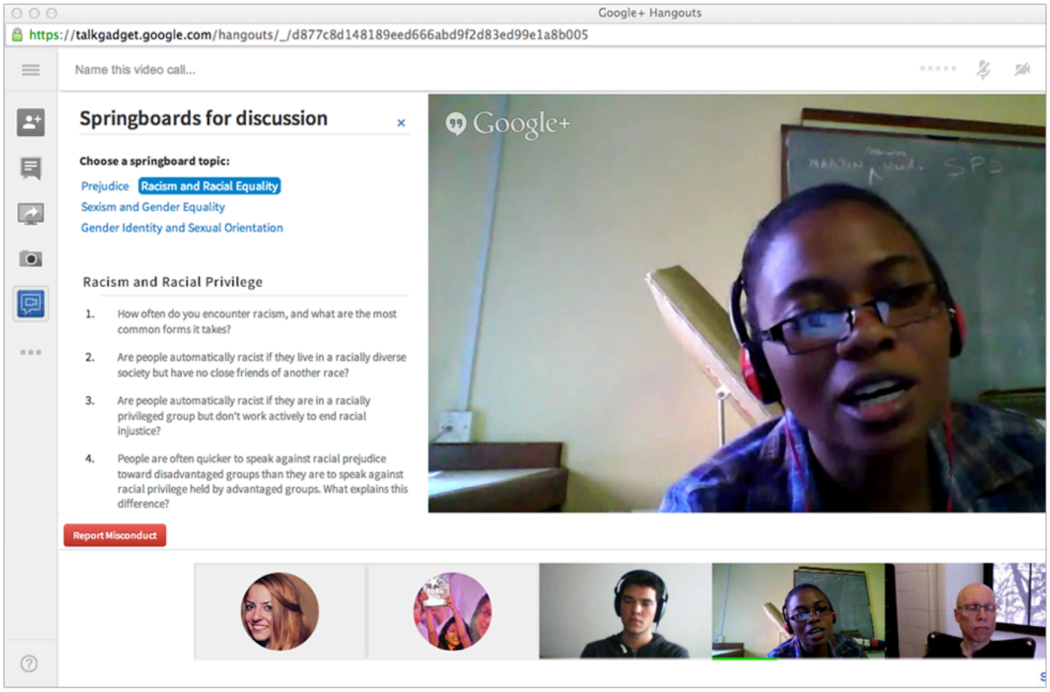
This thesis introduces PeerStudio (www.peerstudio.org), a feedback system built on the insight that in a global classroom, at most times of day, at least a few students are online. This enables students to solicit and receive fast feedback at any time of day, thereby creating opportunities for revision. Unlike our prior work with Coursera, PeerStudio focuses on formative feedback before an assignment is due. Students can submit an assignment draft for feedback at any time. By trading their evaluation for two peers' drafts, students receive rubric-based feedback on their own draft. Students can repeat this process as often as necessary. In a typical online class, the median student using PeerStudio receives feedback within just twenty minutes of submission, and 90% of students receive feedback within an hour.

PeerStudio's key insight is to leverage the temporal overlap between students: as classrooms scale, the expected time delay between requests shrinks to the point of being brief and predictable. PeerStudio augments this natural overlap by selectively emailing students to recruit raters if enough aren't already online. (In practice, around 28% of reviewers are recruited via email.) A controlled experiment in a MOOC found that students who received fast feedback wrote better final essays than students who received feedback delayed by 24 hours, or those who didn't get early feedback at all [19]. Two MOOCs, and several in-person classes in three universities have now deployed PeerStudio to provide students fast in-progress feedback. These deployments have also enabled us to refine its design at multiple scales. Future work could also use similar deployments to understand the causal mechanisms that lead to these improvements.

How can we enrich classroom discussion with massive scale diversity?

Could structured peer interactions enable classrooms to leverage global diversity? And could programmatic control of these interactions encourage students to overcome potential homophily that prevents them from interacting with peers unlike themselves? What if students talked to their global peers in real-time, and learned first-hand from their global perspectives? Could class discussions in a human rights class then become an actual "mini United Nations," as one of our users called it, with students discussing how their countries' policies affect their lives?

Talkabout is a small-group video discussion system that recognizes that the differing experiences and viewpoints of global classmates can help students understand complex topics more deeply.


Google+ Hangouts
https://talkgadget.google.com/hangouts/_/d877c8d148189eed666abd9f2d83ed99e1a8b005
Name this video call...
Google+

Springboards for discussion

Choose a springboard topic:

Prejudice | Racism and Racial Equality
Sexism and Gender Equality
Gender Identity and Sexual Orientation

Racism and Racial Privilege

1. How often do you encounter racism, and what are the most common forms it takes?
2. Are people automatically racist if they live in a racially diverse society but have no close friends of another race?
3. Are people automatically racist if they are in a racially privileged group but don't work actively to end racial injustice?
4. People are often quicker to speak against racial prejudice toward disadvantaged groups than they are to speak against racial privilege held by advantaged groups. What explains this difference?

Figure 6: Talkabout, our global small-group discussion platform leverages the geographic diversity in MOOCs.

The Talkabout system uses Google Hangouts to host globally distributed discussions. Talkabout creates several discussion times that students can choose from. When students arrive, Talkabout groups them with a small number of global peers. Instructors can guide and structure these discussions with a script that Talkabout embeds in every discussion. Our work also developed strategies to create discussion scripts that better utilize diversity. For instance, scripts that ask students to discuss how they related course concepts to their everyday lives improve retention of concepts, and allow them to contrast their thinking with that of peers around the world.

Over 5,000 students from 135 countries and fifteen MOOCs have used Talkabout to discuss topics as varied as organizational behavior, psychology, philanthropy, women's health rights, creativity, and designing effective experiments. Talkabout discussions reflect the global diversity of online classes: the median pairwise distance between participants in discussions is 4,100 miles, approximately the distance between New York and St. Petersburg, Russia.

An experiment with 3,500 students in two massive classes found that geographically diverse group discussions are pedagogically valuable. In both classes, students that participated in more geographically distributed discussions had larger improvements in future grades. Such improvements are likely because talking with diverse peers shifts students from automatic thinking to more active, effortful, conscious thinking.

Unintended effects in global scale peer systems

The preceding chapters describe how software and pedagogy can be co-designed to achieve a desired learning goal. This chapter aims to build theory for learning at scale design praxis by discussing three examples of unintended side effects in our systems. The first is patriotic grading: when we deployed our peer assessment system from Chapter 3 to a global classroom, we found that students rated work from their own country higher than work from other countries. We describe a series of experiments that suggest that this effect may be due to students being unfamiliar with work from distant classmates, or because of an implicit bias that favors work from “in-group” peers. Second, we describe an experimental system we designed to increase students’ commitment to completing assignments. However, we found that the system had the unintended consequence of reducing the fraction of students who completed assignments. We hypothesize that this is because the experimental manipulation crowded out students’ intrinsic motivations with weaker extrinsic motivation. Finally, inspired by project classes where instructors motivate students with examples of excellent student work, we showed students using our peer assessment system examples of excellent work from classmates. Contrary to our expectations, a randomized controlled experiment showed that students who saw such examples actually performed worse on future assignments in the class. We hypothesize this is because while our system showed examples, it did not provide the scaffolding instructors did to help students adopt ideas from such examples. This experience suggests that merely designing software without attending to the underlying learning mechanisms may be ineffective or even harmful.

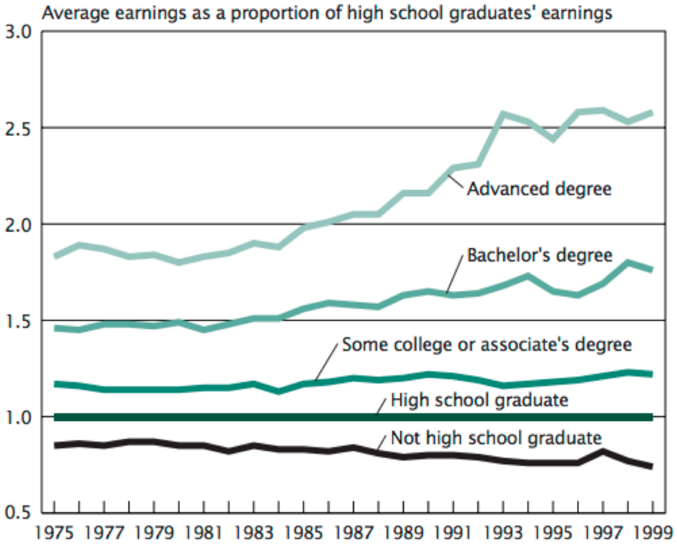
Large-scale validation of social and learning theory

This dissertation is the result of approximately 40 randomized controlled experiments in large online classes, with the average experiment enlisting just under a thousand students. These experiments create opportunities to validate and extend learning theory “in the wild”.

Table 1: Controlled experiments in this dissertation. Average N=1302.

Research question	**Result**
Student engagement	
Do checklists help students complete assignments?	Yes
Does showing growth-mindset based messages improve engagement?	No (ns)
Does sending students personalized re-engagement messages improve engagement?	No (ns)

Does showing activity by friends enrolled in the same class improve engagement?	No (ns)
Do more students respond to authority (instructor asked) or explanation of benefits (for using a new tool)?	Authority
Does reminding students about upcoming deadlines improve retention?	Yes (short-term only)
Does providing students a pro-social reason to help others improve engagement?	Yes
Does receiving a "thank you" note improve student engagement?	Yes
Accuracy in assessment	
Does feedback on accuracy enable students to grade more accurately	Yes
Does asking students to commit to completing work improve engagement?	No, reduces
Do students perceive grades generated from a model that accounts for individual raters' biases and variances to be more accurate?	No, unless told about model
Does explaining a grade correction in greater detail improve students' perceptions of accuracy?	No (ns)
Can students more reliably compute a score provided an interactive checklist instead of free-form prompt?	Yes
Can students verify others' assessment faster than assessing themselves?	Yes
Do students grade peers from other countries lower if submitters are not identified?	Yes
Is students' grading bias affected by linguistic cues?	Yes
Is students' grading bias reduced by better rubrics?	Yes
Discussions and groups	
Does allowing students to participate in discussions improve course outcomes?	Yes
Does geographic diversity in discussions improve course outcomes?	Yes
Does gender balance in discussions affect course outcomes?	No
Can groups more reliably gain critical mass if they require reservations than if available anytime?	Yes
Are ad-hoc groups as successful as groups that allow students to meet with the same group?	Yes
Do assigned moderators improve discussion quality in a Talkabout group?	No
Does showing tips for moderation improve discussion quality?	No (ns)
Does enforcing a discussion script improve discussion quality?	No, reduces
Does showing the results of a "360-degree" discussion review improve engagement?	No, reduces
Does showing aggregate suggestions for improving discussion style improve student engagement?	No, reduces
Improving future learning	



Source: U.S. Census Bureau, Current Population Surveys, March 1976-2000.

Figure 7: Higher education has an increasingly important role in economic success. Compared to those with a high-school diploma, those with bachelor's degrees earn 18% more than they would have in 1975.

Does showing examples of great/approachable work improve students' future work?	Reduces with excellent
Does fast feedback improve grades more than slow feedback?	Yes
Does early feedback improve grades?	Yes

Going forward, this dissertation points to a future where the greatly expanded access afforded by online classes combined with scalable teaching methods can solve large-scale, societal challenges. The greatly expanded access has the potential especially to

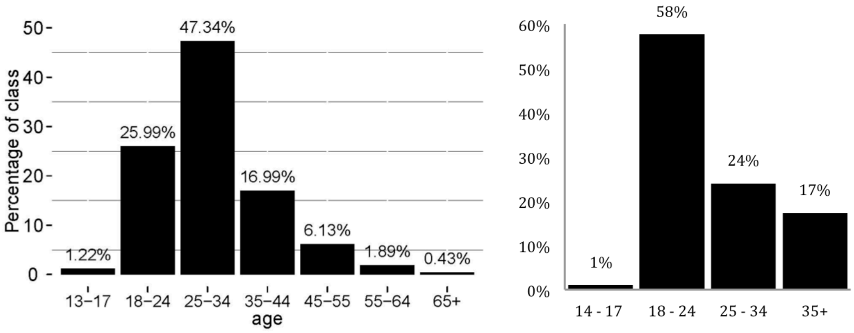


Figure 8: MOOCs attract students across a wider range of ages. Left: enrollment in a typical MOOC (HCI, Fall 2012). Right: Age of college enrollment in the United States, 2012; from [293]**.**

benefit students outside the traditional college demographic. For example, already the median learner in online classes is much older, and is more likely fully employed (see Figure 8 and Chapter 3).

MOOCs could also expand access to higher education beyond the “college demographic” that most universities cater to. Universities are designed for students with minimal employment experience, little familial responsibility, predominantly work full-time on their education [20]. And while the skills and knowledge that students acquire in college lead to their social growth, make them more civic-minded [21], and play an increasingly important role in their overall economic success (see Figure 7), 40% of all high-school graduates in the US still don’t enroll in college [22]. As the demand for specialized, higher education continues to rise [23], massive online classes could allow students to acquire these skills later in life.

Chapter 2
Related work

This chapter provides an overview of related work in education, social computing, and psychology that informs the design of systems introduced in this thesis. This prior work clarifies the benefits that peer interactions bring to the classroom, and suggests the mechanisms that lead to these benefits. We then consider the barriers to scaling peer interactions to the massive scale of an online class. Finally, we discuss work that describes how online communities are designed, and how that informs the design of our systems. While this chapter provides an overview of related research; additional research that informs the design of particular systems is discussed in chapters dedicated to those systems.

The classroom benefits of peer interactions

In the physical classroom, peer interactions have profound benefits. In fact, when students merely believe they are interacting with another person, they recall more of what they are taught, even when the actual interaction may be with a computer agent [24]. The benefits of actual, rich peer interactions are much larger:

Learning benefits of peer assessment, tutoring, and discussion: Peers assessing each other using well-defined criteria result in an audience that provides honest feedback and multiple perspectives [25]. Evaluating peers' work exposes students to solutions, strategies, and insights that they otherwise would likely not see [25], [26], and provides learning gains not seen with external evaluation [27]. Peer assessment also increases student involvement and maturity, and enhances classroom discussion [28]. In addition, it makes classrooms more efficient by lowering the grading burden on staff [9].

Peer tutoring creates an efficient method of instruction, and it also encourages listening and engagement [5]. Peer tutoring can be implemented as, for example, the Jigsaw method [29] of creating “experts” by assigning different sub-topics to each student in a team. These experts then teach other students in the group.

Motivational and metacognitive benefits of peer interactions Peer interactions also improve students’ sense of belonging, and positive interactions with peers (especially in younger students) are correlated with a stronger motivation to engage in pro-social and academic activities [30]. In addition, less competent peers improve communication skills and have more collaborative and more positive social interactions in general [4].

Informal interactions with peers can also lead to metacognitive changes, changing students’ perceived motivations for undertaking activities. For example, students perceiving that their peers expected them to prioritize education and behave pro-socially, participated in academic activities because they thought these activities were enjoyable and important, rather than being required tasks [31].

Given the large benefits of peer interactions, as well as their potential to create a more efficient classroom, practices of collaborative and peer learning are widespread in the physical classroom—81% of US schoolteachers report using it every day [32].

Barriers to scaling peer interactions to an online class

The main barriers to scaling peer interactions to an online class stem from challenges in remotely scaffolding interactions, and in supporting asynchronous, and sometimes hard-to-predict access. In general, the large size of the classroom makes it impossible for instructors to monitor interactions; therefore interactions must be designed so they are beneficial even without close supervision. Furthermore, the large diversity in the class restricts assumptions in how students might participate in these unsupervised interactions. We discuss each of these barriers in turn.

Classroom interactions rely on close supervision

In physical classrooms, peer interactions rely crucially on instructor supervision and guidance. For example, when students informally teach each other, they are likely to lecture to their partner, rarely elaborating on their explanations or allowing their partner to apply information on their own [33], which reduces the effectiveness of their tutoring. Teachers can train students to engage in more productive tutoring behavior [33]. For example, dyadic collaborations improve when one student walks their partner to apply concepts to example problems [34]. It is challenging to scale such close supervision to an online class, as instructors cannot monitor the thousands of simultaneous interactions manually, nor can they intervene.

Students have less predictable engagement online

Many classroom peer interactions are designed assuming all students who are required to participate in an activity will do so. For example, when students are assigned to be "experts" in a jigsaw session, the entire team's learning of the topic is dependent on their participation. However, as Kizilcec and Shneider note [35], students may interact with an online class as they might with a social networking website: "As one of many options on the Web for finding information, socializing, or collaborating, these environments [online classes] are as amenable to casual engagement with content as they are to the focused, ongoing activity characteristic of a student in a traditional course." Therefore, to increase engagement, peer interactions should be designed to either accommodate casual learners, or they must accurately filter for students who will engage as required.

Casual engagement also affects how consistently students engage in the class over time. For example, fewer than 10% of students who are active in an online class during its first week remain active until the last lecture [36], [37]. Many models of classroom peer interactions, such as Student Team Learning [38], suggest that team rewards should be based on improvements in teams' prior performance. However, if the majority of the team does not participate consistently, this reward structure is difficult to operationalize, and may be perceived as unfair, reducing its effectiveness [39]. We discuss how to design online peer interactions for unpredictable participation in Chapter 6.

Students have more diverse motivations online

A common classroom technique for ensuring student participation is making participation necessary for grades, course credits or other credentials. However, only around 45% the students in online classes are motivated by certificates of accomplishment (the equivalent to credentials), about a quarter of students are motivated by the desire to meet new people, while others are motivated by simple curiosity, a chance to improve their language skills, or the prestige of the professor or institution offering the course. A large fraction (56%) are motivated by job prospects [40], which are extrinsic motivations, but less easily operationalized as grades. This means that simple interventions, which offer grades as the only reward may not sufficiently motivate students. We discuss alternate motivation strategies in Chapter 7.

Students are not collocated, participate asynchronously

Most students in MOOCs are not collocated; typically no country represents more than a quarter of participating learners (Chapter 6). One approach to solving the challenges of asynchronous peer interactions at large scale is to break the online classroom into smaller groups that are collocated and synchronous, and then adopt techniques from physical classrooms [41]. While intuitively appealing as a way to reuse the decades of research into developing classroom techniques, this approach still suffers from the problems of minimal supervision, unpredictable engagement, and diverse student motivations. Therefore, this thesis proposes an alternative approach to go “beyond being there” [42]—examine why classroom peer interactions are pedagogically valuable, and then design new online interactions that are suited to the online environment and provide similar benefits.

Scaling peer interactions

This section discusses related research that informs the design of peer interactions in our systems. In designing new peer interactions, we have found research in online communities to be particularly valuable. Recall from the previous section that many students may only engage casually with an online class, similar to how they might use a social network (e.g. Facebook). Furthermore, online communities serve large numbers of members with minimal administrative oversight.

Creating peer interactions that require minimal supervision

If instructors can't monitor interactions closely, could we create norms that lead to productive peer interactions? Descriptive norms (seeing what most people do) are generally more effective than prescriptive norms (such as rules) [43]. In addition, students may consider norms to be fairer than automated corrective action [44]. To create positive descriptive norms about what behavior is acceptable, systems can filter what students see. For example, people are more likely to litter if they saw someone do so in an already littered environment, but *less* likely to do so if they saw someone littering in a clean environment [45].

Creating peer interactions that increase engagement predictability

Recall that participation in online class activities is largely voluntary and difficult to enforce. However, peer interactions could be more predictable if even a fraction of students are committed to participating. Prior work suggests three reasons why people are committed to participating in an online community even when doing so incurs effort, time, or money: affective commitment (because they like the community and identify with it), norm-based commitment (because they think it is the right thing to do), and need-based commitment (they derive a benefit larger than the cost of participation) [46].

These commitment mechanisms suggest that systems can encourage need-based commitment by making the potential educational benefits salient, potentially with the involvement of the instructors. Instructors can also play a pivotal role in creating norm-based commitment, by making peer-learning opportunities a core part of the class (creating a norm of participation). We discuss these strategies in detail in Chapter 7. Finally, they can improve affective commitment. Prior work suggests a particularly powerful way to do so when there is a large turnover in participants as in an online class is to encourage participants to identify with the community as a whole, rather than with particular members of the community [47], [48].

Creating peer interactions that leverage diversity

This dissertation takes the view that diverse motivations and experiences of participants can be a valuable asset. Interactions designed with diversity in mind could use prior work on how workplaces and schools welcome diverse participants. In particular, students benefit most from their classroom's diversity when the numeric representation of diverse groups is large *(structural diversity);* and the number of settings that students interact in is large (*experiential diversity)* [49]. Ideally, students must meet frequently, and with equal status, in situations where collaboration is necessary and stereotypes are disconfirmed [50], and where differing views are welcomed [51].

Together, this prior work suggests that it may be possible to design new peer interactions in online classes in which a large number of diverse students are committed to participate. It also suggests that the design of interactions may not mirror interactions in physical classrooms. To design interactions that still provide the benefits of classroom peer interactions, we turn to *how* classroom interactions yield pedagogical benefits next.

Mechanisms for the classroom benefits of peer interactions

Why do peer interactions improve learning? If systems can be designed with these mechanisms in mind, they could more effectively scale these learning benefits. Prior work provides evidence for three causal paths that explain the benefits of peer interactions: motivation, social cohesion, and cognition. Each causal path has strong supporting evidence, and recent work suggests that all three pathways may work synergistically [5].

Motivation: Peer interactions are an effective way to improve motivation. Some early work went so far as to posit that the entire benefit of peer interactions was motivational [38]. To improve motivation, interactions may be structured to have team goals and reward structures that encourage pro-social contributions. For example, the Student Team Learning model [38] suggests that tasks should have a goal that requires contributions from every member for success, such as making each student's grade the mean of grades of all students in their group. Within this extrinsic reward framework, it also suggests that teams are rewarded for improvements over their own past performance, so even the weakest teams can be rewarded for their effort.

Social cohesion/interdependence: Peer interactions may help students learn from each other in groups by making them like the group members, and by tying their own self-identity to group membership [52]. This suggests that interactions that enable students to develop a stronger sense of group belonging will improve the benefits of peer learning. For example, by creating "experts" who are assigned different sub-topics, the Jigsaw method creates a sense of interdependence in a team, leading to strong cohesion.

Cognition: Peer interactions cause cognitive changes that enable learning. In adults, cognitive changes are primarily a result of elaboration [53]. Peer interactions enable elaboration, the process by which learners create more detailed representations of concepts in memory. Elaboration improves the recall of concepts [54], [55]. Amongst peer interactions, the largest gains in elaboration are through explaining to a peer, though listening to elaborated peer explanations help as well [56]. To encourage elaboration, teachers might ask students to take turns, with one student describing a concept or making an argument, and the other asking for and interpreting evidence [57].

These Causal Pathways Interact: To achieve their potential benefits, peer interactions that are designed to primarily employ one pathway need to also pay attention to others. For example, in controlled lab studies, jigsaw teams perform well without any extrinsic motivation [6]. However, in classroom studies (with extrinsic motivations like grades), jigsaw methods only increase learning if they incentivize students to learn collaboratively, for example by awarding every student in a team the average grade [5].

Together, this prior work suggests that it may be possible to design new peer interactions that capture the benefits of popular small-classroom interactions, and these interactions may succeed if they target student motivation, cohesion and cognition.

A final question we consider in this chapter is how these scalable peer interactions might compare with other models of scalable education, such as modeling student learning computationally and providing automated tutoring assistance, and improving how students learn in isolation.

The relation between peer interactions and other computational learning support

Are the cognitive benefits of peer interactions merely a solution for domains where computational systems that more explicitly model learning and provide students automated support are still unavailable? Kumar and Rose find that on the contrary, peer support and automated tutoring systems can be mutually reinforcing [58]. Having students discuss their homework problems in a structured group can approach the effectiveness of providing individual students with sophisticated computer-generated hints using intelligent tutoring systems. Furthermore, systems that provide hints for how to discuss productively can improve the effectiveness of these interactions even further, without a sophisticated understanding of the domain students are learning.

These results suggest that peer interactions will valuable even when automated support is present. They also present a promising alternative at large scale. Even if systems cannot fully model student learning, they could still find peers with just the right levels of understanding and provide them support to discuss productively [59].

The relation between peer interactions and data-enriched solitary learning

Large classes can also employ data-driven systems that improve a student's solitary learning experience. Unlike the focus of this dissertation, these techniques focus on improving the experience of a student using class resources alone, and are largely complementary to the peer-interactions presented in this dissertation. In general, these techniques leverage redundancy in how students interact with class resources.

Redundancy enables interfaces that summarize and visualize student behavior. For example, students re-watch parts of videos that are confusing, important, or relevant to questions on a test more frequently. Furthermore, students from countries with smaller classrooms watch videos non-linearly more frequently [60]. Lectures cape introduces techniques that use these observations to create visualizations of student interest and confusion [60]. Similarly, Over Code introduces techniques for visualizing students' programming solutions [61]. Clustering student submissions enables instructors to see common student errors, and enables them to comment on a large number of solutions at once [61]–[63].

This thesis leverages redundancy to improve peer interactions. Redundancy in student submissions to short-answer questions is used to reduce the number of peer raters who assess it; more raters are assigned to a novel response, fewer to responses that are similar to many others (Chapter 4).

Future work could also leverage redundancy more directly in peer communication. For example, a system could synthesize a student discussion in a quasi-interactive manner, with the student's responses replaying common fragments of previous student discussions on the topic. Perhaps such simulated conversation could provide some of the benefits of a real discussion [24], while a student waits for the system to find a real conversation partner.

Chapter 3
Peer assessment in massive online classes[1]

Over the past few years, more than a hundred thousand students have earned certificates in large online classes—on topics from Databases to Sociology to World Music—and millions have signed up [64]. These MOOCs provide students on-demand video lectures, often along with automated quizzes and homework, and class forums that allow students to interact with each other.

Many such classes use automated assessment (e.g. [65]), which precludes the open-ended work that is a hallmark of education in creative fields like design [66]. Furthermore, viewing and critiquing others' work plays a key pedagogical role in these domains [11]. Fields like design have also traditionally relied on intimate co-location to enable these activities and to confer values and norms [11]. However, in a global, online classroom, students lack the shared context co-location provides. How can we scale both evaluation and peer learning in creative domains online?

One approach for scaling assessment and peer learning would be for students to evaluate their peers' work. Peer assessment potentially enables large classes to offer assignments that are impractical to grade automatically. Furthermore, human grading more easily provides context-appropriate responses and better handles ill-specified constraints [67]. But, how can students who are novices themselves be motivated and trained to perform peer assessment well? This chapter reports on our experiences with the first use of peer assessment in a massive online class. It is the largest use of peer

[1] A version of this chapter was originally published as an article in the ACM Transactions of Computer Human Interaction as [37].

assessment to date. As of June 2013, this technique has since been adopted in many other classes, including 79 MOOCs on the Coursera[2] platform alone.

The design studio as an inspiration

For over a century, the studio has been a dominant model for architecture and design education, and has expanded into fields including product design [68], HCI [69], [70], and software design [71]. This chapter considers the studio as an inspiration for online design education.

The studio model of education was formalized in the Cole de Beaux-Arts [72]. Studios provide an open, shared environment for students to work. This co-presence provides social motivation and facilitates peer learning through visibility of work [73]. Formal and informal studio critique helps students iteratively improve their work [11].

Public visibility of self and peer work provides students with a nuanced understanding of design. In particular, seeing their peers' work along with their own work through its evolution allows students to understand decisions and tradeoffs both in their own designs, and in those of their peers [25].

Formative studio feedback further engages students in reflective practice [11]. Informal, formative feedback is often through oral critiques or "crates" by teachers or other experts [74]. Such informal, qualitative feedback is essential, because it encourages iterative practice [75]. Because crates are often delivered in public, students also learn from observing peer work as well as by working on their own [76].

Expert critiques also serve as summative assessment. Experts often assess design based on trained but tacit criteria [77]. Amiable et al demonstrate that expert consensus is a reliable measure of the quality of creative work [78]. Their Consensual As-

[2] https://www.coursera.org/

sessment Technique asks experts to rate artifacts on a scale, and provides no rubrics and does not ask raters to justify their rating. Other techniques provide an assessment process to observe, interpret and evaluate work [79].

The design studio suggests three requirements for successful design education online. First, it must support open-ended design work with multiple correct solutions. Such work is especially important in design education because successful design often requires generating and reflecting on multiple ideas [66], [80], and on exploration and iteration [81]. Second, assessment must allow students to learn the tacit criteria of good design. Criteria for good design are often not explicitly defined [82]. For instance, interactive interfaces may be subjectively evaluated for whether they are learnable and appropriate [83], criteria that require tacit interpretation. Third, assessment must provide students both qualitative formative feedback, and summative feedback.

The promise of peer assessment

Open-ended assignments generally rely on human graders. The inherent variability of open-ended solutions and lack of defined evaluation criteria for design makes automatically assessing open-ended work challenging [84]. In addition, automated systems frequently cannot capture the semantic meaning of answers, which limits the feedback that they can provide to help students improve [67], [85].

The time-intensive, personalized assessment of grading sketches, designs, and other open-ended assignments requires a small student-to-grader ratio [86], [87]. This staff effort is prohibitive for large classes: staff grading simply doesn’t scale.

Peer and self-assessment is a promising alternative. It not only provides grades, it also importantly helps students see work from an assessor’s perspective. Peer feedback in design classes also creates an audience that provides honest feedback and multiple perspectives [25]. Evaluating peers’ work also exposes students to solutions, strategies, and insights that they otherwise would likely not see [25], [26]. Similarly, self-assessment helps students reflect on gaps in their understanding, making them more resourceful, confident, and higher achievers [30], [88], [89] and provides learning gains not seen with external evaluation [27].

Peer assessment can increase student involvement and maturity, lower the grading burden on staff, and enhance classroom discussion [28]. Peer assessment has been used in co-located classroom settings for many different kinds of assignments [90], including design [25], [91], programming [26] and essays [92]. How can we make this classroom technique scale to a large online class?

Scaling peer assessment

In-class peers can assess each other well [93]–[95], suggesting the viability of the technique, at least with in-person training. To effectively scale peer assessment, we can learn several lessons from crowdsourcing [96]. First, crowd workers perform better when they are intrinsically motivated by the task's importance [97]. Second, consensus among raters serves as a useful indicator of quality [98]. Third, interfaces like FoldIt [13] and NASA Clickworkers [99] demonstrate that short, well-crafted training exercises can enable legions of motivated amateurs to perform work previously thought to require years of training. These peer-sourced systems introduce new challenges and opportunities beyond crowd sourcing. For example, students using peer assessment both create the work to be assessed and perform the assessment. One theme this chapter will explore is the learning benefits that arise from those dual roles.

Massive online classes provide a valuable living lab [100], [101] for exploring peer-sourcing approaches, and our hope is that peer-sourcing insights from massive classes will contribute techniques that apply more broadly (this has since come true, as we shall see in later chapters).

Contributions

This chapter reports on our experiences with peer assessment over two iterations in the first large-scale class to use it (http://www.hci-class.org). Since our adaptation of peer assessment to MOOCs, variations of the system described here have since been used in dozens of other large online classes, including Mathematical Thinking, Programming Python, Listening to World Music, Fantasy and Science Fiction, and Sociology.

Over both iterations of the class, 5876 students submitted at least one assignment and participated in peer assessment. Overall, the correlation between peer grades and staff assigned grade was r = 0.73, and the average absolute difference between peer and staff grades was 3% (positive and negative errors were approximately balanced).

In end-of-course surveys, students reported both receiving peer feedback and performing peer assessment to be valuable learning experiences. On a seven-point Likert scale, the median rating was 6 (7=very valuable). Surprisingly, 20% of students voluntarily assessed more submissions than required.

We explored several techniques to improve assessment accuracy and encourage qualitative feedback. First, we found that giving students feedback about whether they scored peers high or low increased their subsequent accuracy. A between-subjects experiment found a 0.97% decrease in mean error (6.77% in the experimental group, vs. 7.74% in the control group). Second, to help students provide peers with high-quality personalized feedback, we introduce short, customizable feedback snippets that address common issues with assignments. 67% of students obtained open-ended peer feedback using this method. Third, we introduce a data-driven approach for improving rubric descriptions. We distinguish items with high student: staff correlation from those with low correlation, and observed the ways they differ to improve the low-correlation ones. After making these changes, the mean error on grades decreased from 12.4% to 9.9%.

The educational environment of a massive online class

This online class is an introduction to human-centered interaction design. The class is offered free of charge, and is open to any interested student. Material covered in class is based on an introductory HCI course at Stanford University. Over the class duration, students watch lectures, answer short quizzes and complete weekly assignments. In a typical week, students watch four videos of 12-15 min each. Videos total approximately 450 minutes across the class, and contain embedded multiple-choice questions.

Multiple choice quizzes tested students' knowledge of material covered in videos. Most significantly, students completed five design assignments. Each assignment covered a step in a course-long design project where students design a Web site inspired by one of three design briefs (Figure 11).

Students who complete the course with an average assignment score of 80% or above earn an electronic "Statement of Achievement" for a Studio track (but no university credit). 501 students earned this statement in the first iteration, and 595 did in the second. 1,573 received a statement of achievement for the Apprentice track comprising watching videos and quiz performance in the first iteration, and 1,923 did in the second.

By the numbers

Similar to other online classes [102], the online HCI class attracted numerous and diverse participants. 30,630 students watched videos in the first iteration, and 35,081 did in the second (32.5% of students in each iteration were female). 55% of students reported they had full time jobs (in both iterations). The median age range in both iterations was 25-34, with a broad spread (Figure 9). In both iterations, students from 124 countries registered for the class and roughly 71% were from outside the United States. Students transcribed lectures in 13 languages: English, Spanish, Brazilian Portuguese, Russian, Bulgarian, Japanese, Korean, Slovak, Vietnamese, Chinese (Simplified), Chinese (Traditional), Persian, and Catalan.

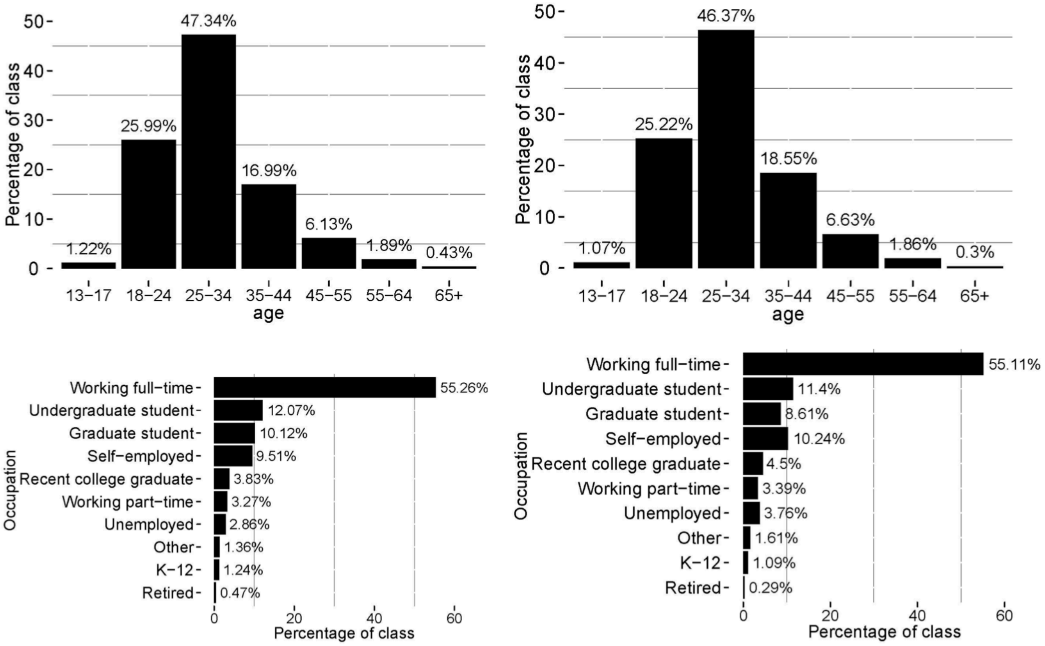


Figure 9: Online classes attract students who cannot use traditional universities, such as those working fulltime. The age distribution of the class is remarkably similar across both iterations. (a) Spring 2012 (iteration 1), 10,190 participants, (b) Fall 2012 (iteration 2), 17,915 participants.

In all, 2,673 students submitted assignments in the first iteration, and 3,203 in the second (Figure 10). The second iteration also allowed students to submit assignments in Spanish; 223 students did so. Student questions were answered exclusively through the online class forum. Across the course, the forum had 1,657 threads in the first iteration, and 2,212 in the second.

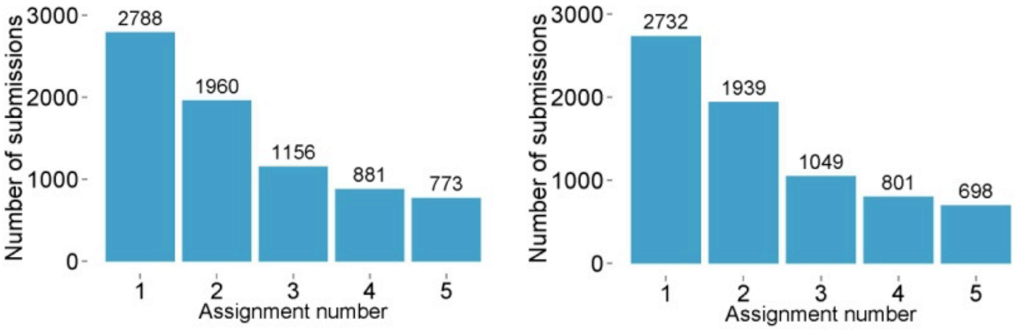


Figure 10: Number of students who submitted each assignment, in iteration 1 on left, iteration 2 on right.

Assignments

All assignments were submitted online, and graded with calibrated peer assessment. Three of the five assignments asked students to create physical artifacts like paper prototypes and upload photographs of their work.

Each assignment included a rubric that described assessment criteria [8]. Rubrics comprised guiding questions or dimensions that student work was graded on, and gradations of quality for each dimension, from poor to excellent. Rubrics were released with the assignment, so students could refer to them while working. Table I shows a part of the rubric for the User Testing assignment, another rubric is shown in Table V.[3]

Peers assessed using the rubric, and students were informed that peers could see all submitted work while grading. Students could also share their peers' work via class forums after grading was complete and staff used examples of student work in class announcements and lectures. Students could optionally mark their submissions as private to prevent such sharing outside the peer assessment system: over both iterations combined, 13.5% of students chose to do so.

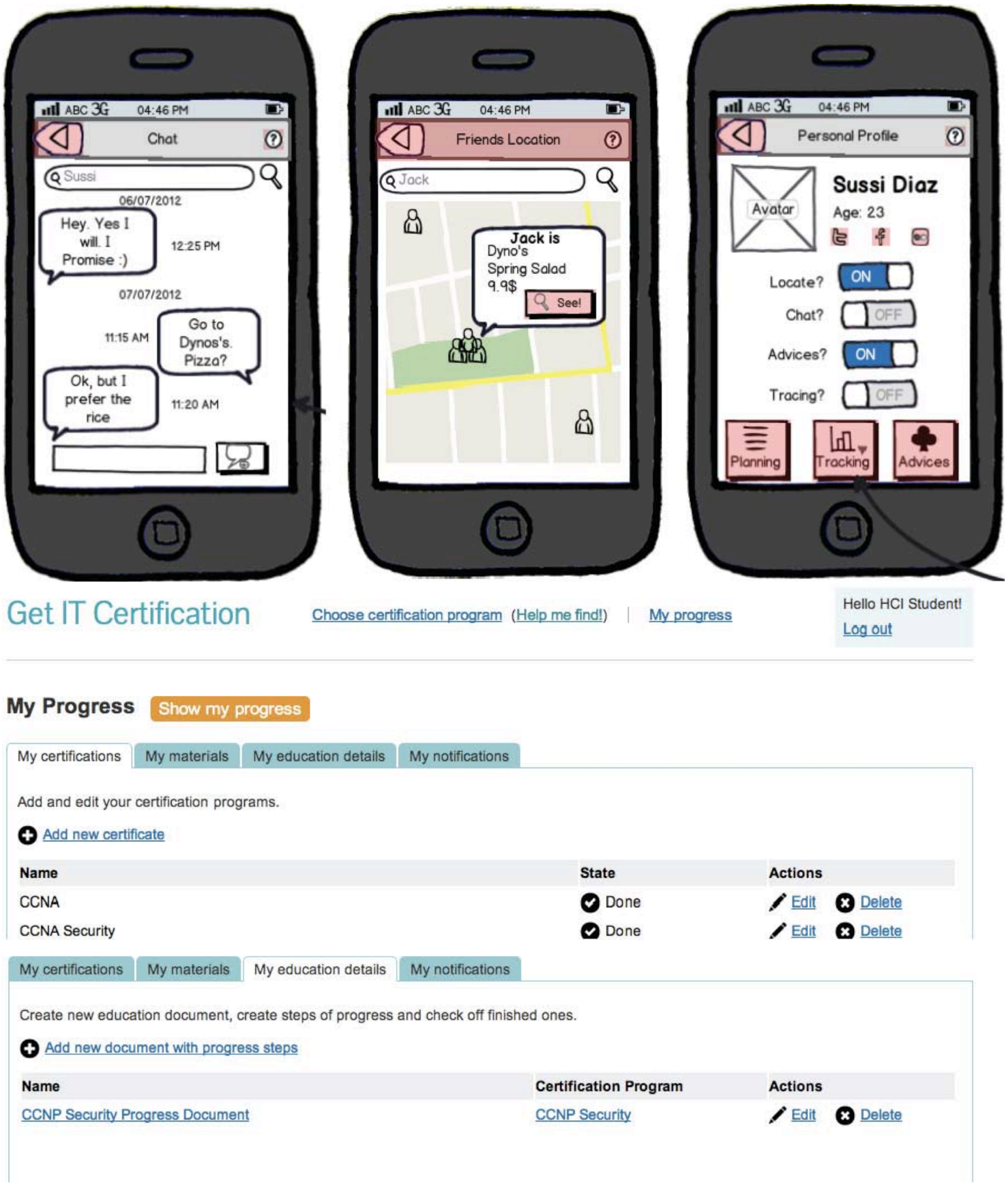
All assignments and rubrics were based on corresponding materials from the introductory HCI class at Stanford.[4] The in-person Stanford class uses self-assessment and staff grading, but not peer assessment.

[3] All assessment materials are also available in full at http://hci.st/assess

[4] https://cs147.stanford.edu/

Peer Assessment

Assessment used Calibrated Peer Review [94]. Calibrated peer review helps students learn to grade by first practicing grading on sample submissions.



Add and edit your certification programs.

Add new certificate

Name	State	Actions
CCNA	Done	Edit Delete
CCNA Security	Done	Edit Delete

Create new education document, create steps of progress and check off finished ones.

Add new document with progress steps

Name	Certification Program	Actions
CCNP Security Progress Document	CCNP Security	Edit Delete

Figure 11: Example prototypes from student projects in the online class (top: early prototype of a social dining app; bottom: a tracker for professional certification at the end of the class.

Immediately after each submission deadline, staff evaluated about a dozen submissions– eight were used to train students; the rest were used to estimate accuracy of assessment. The next day, peer assessment opened for students who submitted assignments. Students had four days to complete peer assessment.

Peer grading for each assignment had two phases: calibration and assessment. During the first, calibration, phase, students see the staff grade for a submission they grade, along with an explanation. If the student and staff grades are close, students move to the assessment phase. Otherwise, students grade another staff-graded assignment. This process is repeated until student and staff grades match closely, with up to five such training assignments. After five submissions, students moved to the assessment phase regardless of how well they matched staff grades.

Then, students assessed five peer submissions. Unbeknownst to the students, one submission was also graded by staff to provide a measure of assessment accuracy. By symmetry, this means that at least four randomly selected raters saw each student's submission, and that each student saw one staff-assessed submission per assignment. Immediately after assessing peers, students assessed their own work. Self-assessment and peer assessment used identical interfaces.

Time spent on assessment varied by assignment. Depending on assignment, 75% of assessments were completed in less than 9.5 minutes to 17.3 minutes. On the median assignment, 75% of assessments took less than 13.1 minutes.

One pedagogical goal of the class was to have students understand and have some influence on their grades. At the same time, we didn't want to reward dishonesty or delusions. To balance these goals, when the self-assessed score and the median peer score differed by less than 5%, the student got the higher score. If the difference was larger, the student received the median peer-assessed score. This policy acknowledges

5% to be a margin of error and gives the student the benefit of doubt. Peer grades were anonymous; students saw all rater-assigned scores, but not raters' identities. Similarly, submitters' names were not shown to raters during assessment, i.e. the assessment system was double blind.

Because assignments built on each other, it was especially important to get timely feedback. Grades and feedback were released four days after the submission deadline (the subsequent assignment was due at least three days after students received feedback). Students who didn't complete either the self assessment or peer assessment by grade-release time were penalized 20% of the assignment grade. Students were allowed to assess more than five submissions if they wanted to (Figure 7 shows the distribution of assessments completed). These additional submissions were also chosen randomly, exactly like the first five submissions.

How accurate was peer assessment?

Methods

To establish a ground-truth comparison of self and staff grades, each assignment included 4 to 10 staff-graded submissions in the peer assessment pool, selected randomly. Across both iterations, staff graded 99 ground-truth submissions. Each student graded at least one ground-truth submission per assignment; a ground-truth assignment had a median of 160 assessments. (Some students graded more than one ground-truth submission per assignment because the system would give them a fresh ground-truth assignment when they logged-out without finishing assessment and returned to the website after a long time).

This chapter's grading procedure assigns the median grade from a small number of randomly selected peers (e.g. 4-5). We evaluated the accuracy of this grading process using the 99 assignments with a staff grade. To simulate the median-grade approach, we randomly sampled (with replacement) five student assessments for each ground-truth submission, and compared the sample's median to the staff grade[5]. We present results for 1,000 samples of five assessments per submission. This sampling method is essentially a bootstrapped statistical analysis [103]. It allows staff to only evaluate a small set of randomly selected submissions, and still provides an estimate for every peer-rater's agreement with their grade (since all peers see at least one staff-graded submission.) Repeatedly sampling five grades from the pool of peer grades provides an approximate distribution of agreement between staff and peer grades.

We also compared students' self grade with their median peer grade to measure whether students rate themselves differently than their peers.

To enable comparisons, we present results for both iterations separately. The second iteration of the course had grading rubrics improved using data from the first iteration (discussed in Section 6.1). The general similarity in accuracy across both iterations (with improvements in the second) suggests that the peer assessment process produces robust results. The second iteration also allowed students to submit assignments in Spanish. For consistency, our analysis does not include those submissions.

At the end of the class, students were invited to participate in a survey; 3,550 students participated in all. Participation was voluntary, students were not compensated, and the survey did not count towards course credit.

Results: Grading agreement

Here, we present percentage differences between peer and staff grades (summarized in Table 2). Most assignments in this class were out of 35 points. Therefore, a 5% difference represents 1.5 points (grades could only be awarded in multiples of half a point).

[5] Staff comprised graduate students from Stanford. The second iteration had Community TAs chosen among top-performing students in the previous iteration in addition to Stanford staff.

For the first iteration, 34.0% of submissions had a median peer grade within 5% of the staff grade, and 56.9% within 10% (Figure 12). The second iteration improved to 42.9% within 5% of the staff grade, and 65.5% within 10%. In the first iteration of the class, 48.2% of samples had a peer median lower than staff grade, 40.2% had it higher. The second iteration had 36% of samples had a peer median lower than staff grade, 46.4% had it higher. Students tended to get better at grading over time (See Section 3.8).

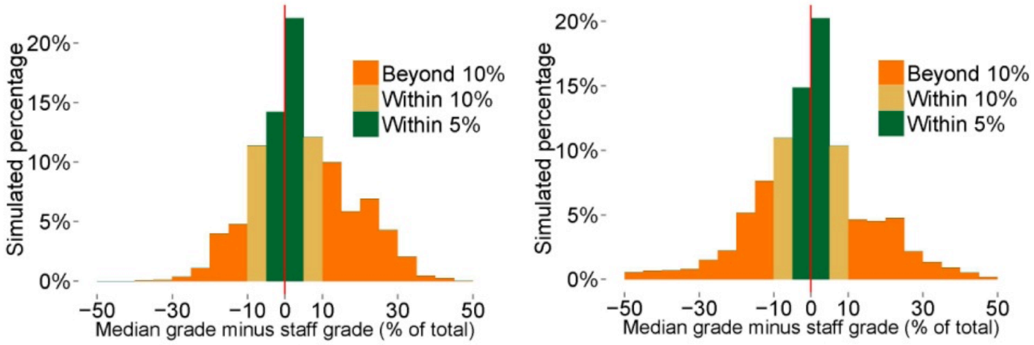


Figure 12: Accuracy of peer assessment for submissions that were graded independently by teaching staff and peer assessors (all five assignments). Graph accuracy of random sample of 5 graders against staff. (left) Iteration 1: 34.0% of samples within 5% of the staff grade, and 56.9% within 10%. (right) Iteration 2:. 42.0% of samples within 5% of the staff grade, and 65% within 10%.

Table 2: Summary of grade agreement. In the second iteration of the class, peer-staff agreement increased, while peer-self agreement decreased.

Metric	Iteration 1	Iteration 2
Peer-staff agreement (within 5%)	34.0%	42.9%
Peer-staff agreement (within 10%)	56.9%	65.5%
Peer < Staff	48.2%	36.0%
Peer > Staff	40.2%	46.4%
Peer-self agreement (within 5%)	28.7%	24.0%
Peer-self agreement (within 10%)	44.9%	40.6%

In the first iteration of the class, 28.7% of submissions had their median peer grade within 5% of the self-assessed grade, and 44.9% within 10% (Figure 13). The median submission had a self-grade 6% higher than the median peer grade. In the second itera-

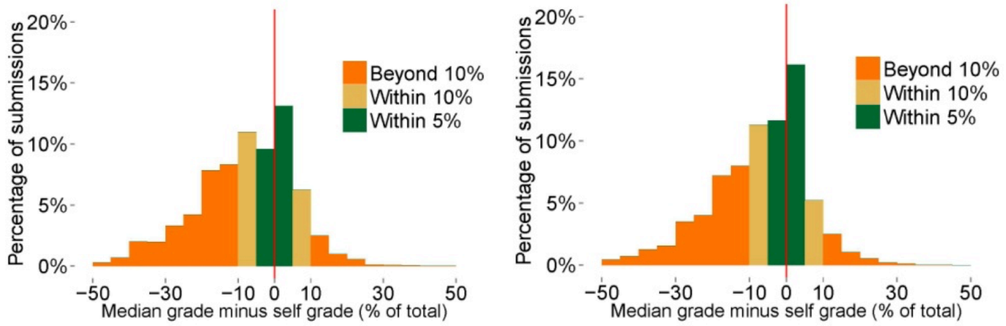


Figure 13: (a) Comparison of median peer grades against self grades. In the first iteration 28.7% of such samples were within 5% of the staff grade, and 44.9% within 10%. (b) Same graph for second iteration of the class. 24.0% of such samples were within 5% of the staff grade, and 40.63% within 10%.

tion, 24.0% of submissions had their median peer grade within 5% of the self-assessed grade, 40.63% had the median peer-grade within 10%. The median submission had a self-grade 7.5% higher than the median peer grade. (We discuss possible reasons for this lowered agreement in Section 6.3.)

Results: Grading agreement between staff

The first two iterations of the class had only one staff member grading each ground-truth submission. To get an idea of how well staff grades agree amongst themselves, in the third iteration of the class we asked multiple staff members to rate each submission.

Submissions were randomly assigned to three staff members (there are six staff members in all). Staff rated 50 submissions over the course.

For these submissions, the average disagreement between staff raters (defined as the median difference between a staff grade, and the mean staff grade) was 6.7%. Of submissions, 28% had all staff grades within 5% of the assignment grade, and 42% within 10%. In contrast, over the second iteration of the class, the average disagreement between peer raters was 25.0%. Only 4.0% of submissions had all peer grades agreeing within 5%, and 16.9% within 10%.

These results suggest that correlation amongst staff grades is many times higher than agreement amongst peer raters. They also suggest that aggregating peer grades leads to a remarkable increase in agreement with staff grades (Section 3.2).

Staff differences in grading were usually due to differing judgments or interpretation. For example, an early assignment asked students to create storyboards of user needs without constraining to a particular design. Staff members differed in how constraining they thought storyboards were.

Such differences suggest the inherent limitations of independent assessment via rubrics due to differences in judgment. Consensus-based mechanisms that encourage sharing perspectives may improve agreement [78].

Comparison to in-person classes

These accuracy numbers also compare well to accuracy in in-person classes. The Fall 2012 version of the in-person class (CS 147) that this class is based on used self-assessment, but not peer assessment. The in-person class had 32.8% of submissions with a self-grade within 5% of staff grade, and 60.8% of submissions within 10% (Figure 14).

Results: Student reactions

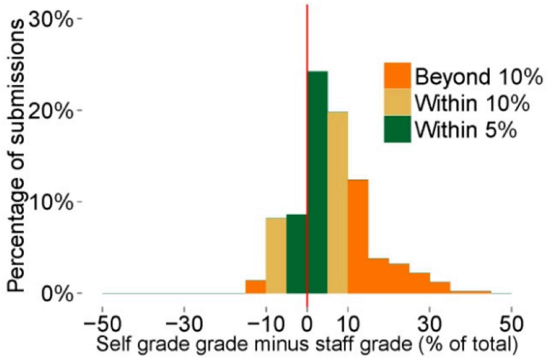


Figure 14: Agreement of self and staff grades in an in-person class.

Table 3: The most frequent trigrams (three word phrases) in students' self-report (over both iterations of class): Students reported both peer and self assessment to be valuable for different reasons.

to see other	114	my own work	175
how other people		your own work	
see how other(s)			
other's work/other people's			
point(s) of view	36	compare my work	50
		I could compare	
compare	12	I didn't/did not	31
helped me understand	12	what I did	19
		point of view	15
"In what ways was assessing others' work useful?" Students frequently cited inspiration, finding example work to critique, and seeing different points of view.		**"In what ways was assessing your own work useful?"** Students frequently cited new perspectives on revisiting work, comparing work to peers', and better identifying their mistakes	

Student reactions to the peer assessment system were generally positive, and 20% of students completed more peer assessments than the class required them to (Figure 15). We infer from this that students found rating their peers valuable or enjoyable, and/or they believed it would help their peers.

Of all students, 42% cited seeing other students' work as the biggest benefit of peer assessment, 31% reported learning how to communicate their ideas as a benefit. Students reported both self assessment and peer assessment to be valuable, and that they played different roles. Evaluating peers was useful for inspiration and to see other perspectives. Self assessment provided students an opportunity to look at their own work again, and encouraged comparing it with others' work they had assessed. It was also useful for identifying mistakes and reflection (Table 3). Overall, students reported learning more by assessing their peers than by assessing themselves: mean ratings were 4.97 and 4.51 respectively for peer and self assessment (6-point Likert scale, 6: "agree strongly (sufficient effort)"), on a Mann-Whitney U-test (U = 580, 562, $p < 0.001$.)

However, students also reported that they felt their peers put in less effort into peer assessment than they did (Table 4). On a Mann-Whitney U-test, mean ratings were 4.57 for peer-effort and 5.46 for their own effort (6-point Likert scale, 6: “learnt a lot”), U = 610, 728, p < 0.001. Reasons for this bias are probably similar to the illusory superiority effect [104]. Designing peer assessment interfaces that emphasize reciprocity and minimize this bias remains future work.

Does a different weighting of peer grades help?

Using the median of peer grades is simple, easily explainable, and robust to outliers. Would a different weighting of peer grades more accurately mimic staff grades?

Method: To find the best linear combination of weights, we built a linear regression on the staff grade with five peer grades in increasing order as the predictors, and with no intercept. This regression seeks weights on peer grades that maximally predict the staff grade.

Results: The best linear regression doesn’t materially improve accuracy. The linear model weighted the five peer grades from lowest to highest at 15.6%, 13.6%, 21.3%, 27.6%, 18.3%. Holding out 10% of ground truth grades, and testing on samples drawn from them, the regression model yields an accuracy of 35.8% of samples within 5%, and 58.8% within 10%. In contrast, using the median yields an accuracy of 35% of

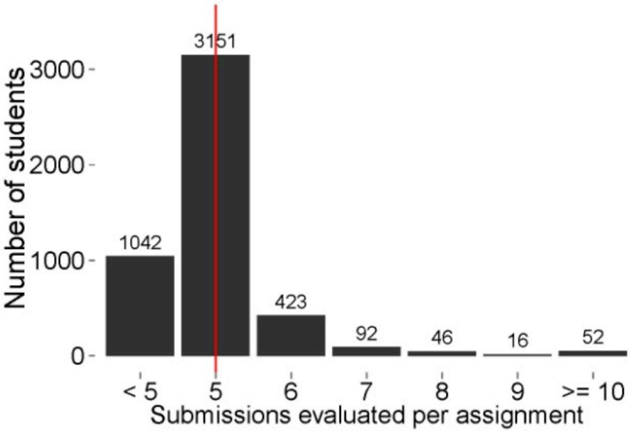


Figure 15: Average number of submissions assessed per assignment (both iterations). Students were required to assess five, and 20% of students evaluated more than required.

samples within 5%, and 58.7% within 10%.

Similarly, the arithmetic mean, geometric mean, and a clipped arithmetic mean (that only considers the middle three grades) all do worse than the median. In addition, errors are approximately evenly spread across the median, so adding a constant correction term to the median grade does not significantly improve accuracy either.

In summary, the simple median strategy seems to be surprisingly effective at identifying the most plausible grade. Is this accuracy sufficient? For a class with letter grades, greater accuracy is needed (because currently about 40% of assignments are a full letter grade away). However, a student's grade for the entire course is generally more accurate due to positive and negative errors canceling out. Using repeated sampling, we estimate more than 75% of students got a course grade within 5% of staff grade (assuming grades in different assignments are uncorrelated). Consequently, for a pass/fail class (such as many current MOOCs, including ours), this accuracy is sufficient for the vast majority of students. We estimate that fewer than 45 students (approx. 6%) were affected by grading errors in each iteration of the class.

Would more raters help?

Increasing the number of raters per submission helps accuracy, but quickly yields diminishing returns (Figure 16). A large number of students rated staff-graded assign-

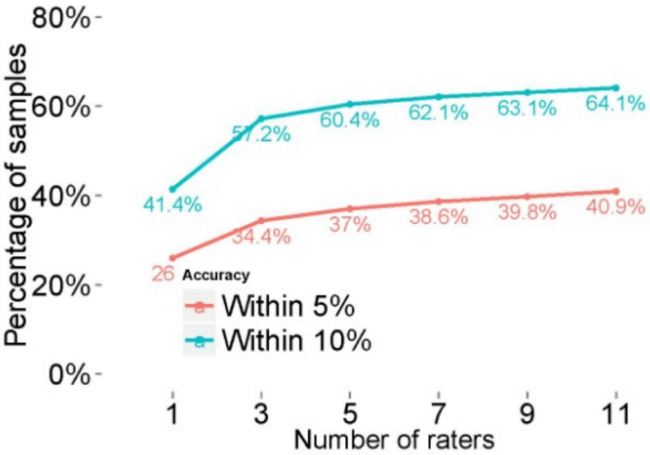


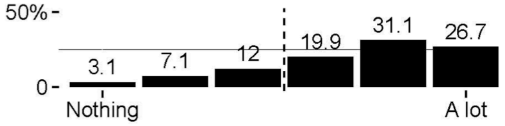
Figure 16: Increasing the number of raters quickly yields diminishing returns.

ments. These allow us to simulate the effect of having more raters. Increasing the number of assessments per submission from 5 to 11 increases the number of assignments that were graded within 5% of the staff grade by 3.8%, and those graded within 10% by 3.6%. Increasing the number of assessments to an (unreasonable) 101 per submission increases the number of submissions graded within 10% of the staff grade by 8.1%.

Table 4: End course survey results (n=3,550) about student perceptions on peer assessment. Students reported learning from assessing others' work than their own, and putting effort into grading fairly.

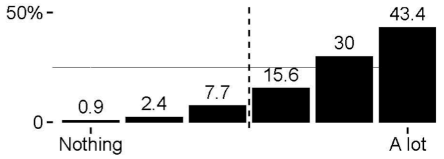
Learned from assessing your own work?

(Nothing … A lot)

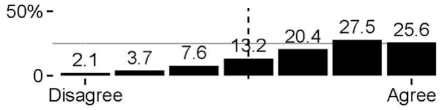


Learned from assessing others' work?

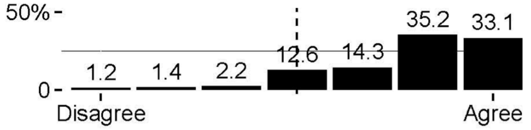
(Nothing … A lot)



The peer assessment process was easy to understand (Disagree … Agree)

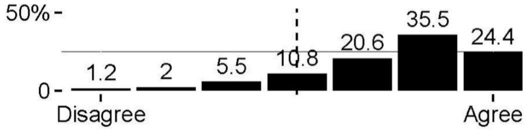


I assessed myself fairly and accurately (Disagree… Agree)



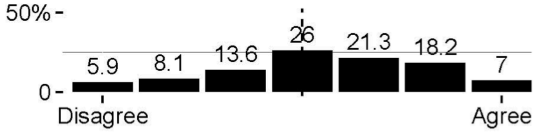
I put sufficient effort into grading peers

(Disagree. . . Agree)



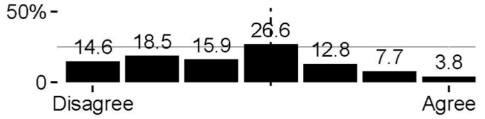
Peers put sufficient effort into grading me

(Disagree. . . Agree)



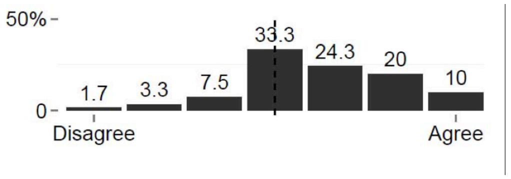
My peer graders did not understand my work

(Disagree. . . Agree)



Rubrics helped me understand exactly what assignments required me to do

(Disagree. . . Agree)



Do students become better graders over time?

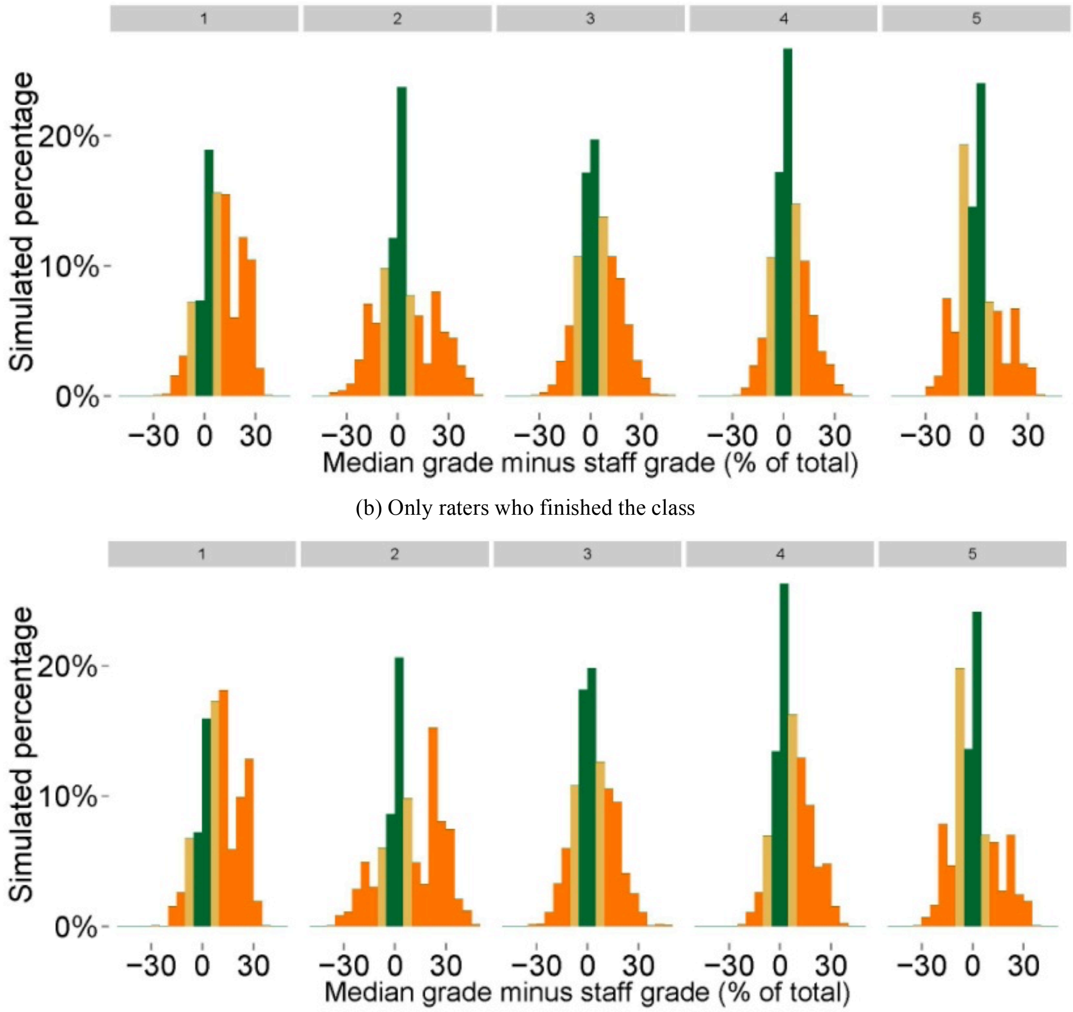
Agreement of peer grades with staff grades generally increases across the class. This increase is seen both for the class as a whole, and for students who submit all assignments, i.e. excluding students that drop out. This suggests that, regardless of individual differences in perseverance and motivation, familiarity and practice with peer assessment leads to more accurate assessments.

Using the repeated sampling scheme described in Section 3.1, five assignments had 26.4%, 36.2%, 36.9%, 43.9%, and 36.8% of submissions estimated within 5% of the staff grade. Within a 10% range, the assignments had respectively 49.1%, 53.6%, 60.9%, 68.5%, and 64.3% within 10% (Figure 17(a)). If we only consider raters that finished the class (and exclude those that dropped out), we see that staff agreement increases as well. The five assignments in order had 23.7%, 29.4%, 38.4%, 39.5%, 37.1% within 5% of staff, and 47.4%, 63.8%, 61.8%, 63.3%, 64.2% (Figure 17(b)). Note that both these numbers are based on repeated sampling from a smaller number of staff-graded assignments. As such, they are more susceptible to variations in staff grades for a particular submission.

What is the right granularity of grades?

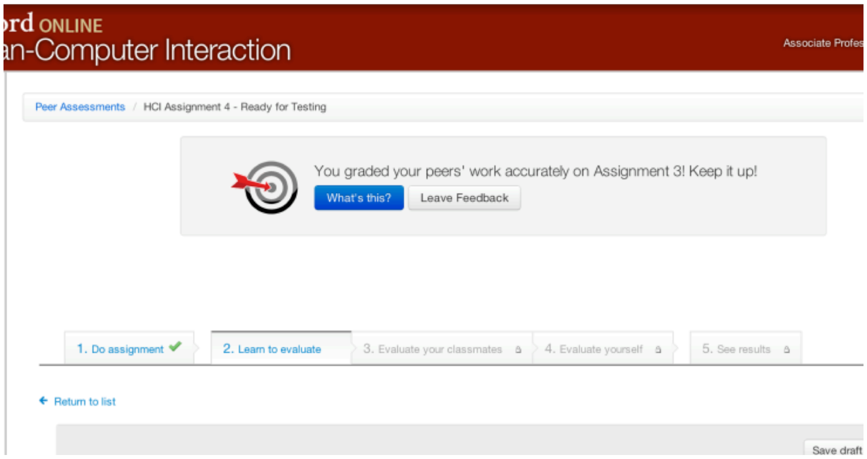
The previous sections show that the grading agreement between staff members, and between staff and students in an in-person class are similar. These differences may approximately represent the smallest discernible differences in quality.

Recall that a 5% difference in grades is 1.5 points in a 35 point assignment, i.e., three times a “just-noticeable” difference in quality (0.5 points, the minimum granularity of grades). Indeed, the in-person version of the class adopted the current 35 point grading scheme (replacing its 100 point scheme from prior years) to better balance accuracy with meaningful differences in quality.


(a) All raters

(b) Only raters who finished the class

Figure 17: Agreement of median peer grades and staff grades across different assignments. (These agreement distributions are more susceptible to variations in staff grades for a particular submission because they are based on repeated sampling from a smaller number of staff-graded assignments.)



You graded your peers' work accurately on Assignment 3! Keep it up!

Figure 18: Students improved grading when provided accuracy feedback. (Screenshot shows feedback displayed when the raters' grade agree well with staff grades.)

Providing students feedback on grading accuracy improves their subsequent performance

So far, this chapter has characterized the accuracy of large-scale calibrated peer assessment. This section explores a feedback intervention to improve graders' accuracy. Prior work has demonstrated that feedback improves the quality of crowd work [105], but can it help raters overcome their (possibly unintentional) grading bias? This section describes an experiment that provided students feedback whether they were grading either "too high," "too low," or "just right," based on how well their grade agreed with staff grades for the previous assignment. We hypothesized that providing students grading feedback would help improve accuracy. We conducted a controlled experiment on the course website that measured the impact of this feedback on accuracy.

Participants and setup

We randomly sampled 756 participants from students who had completed the second assignment of the second iteration of the class.

The between-subjects experimental setup had two conditions: a no-feedback control condition where students received no feedback on the accuracy of their grading, and a feedback condition that provided feedback on their grading bias: too high, too low, or just right (Figure 20).

To generate bias feedback, the system compared the participant's rating and the staff rating of the previous assignment's ground-truth submission.

If the rating differed by more than 10%, then feedback was shown as too high/too low; otherwise the feedback was "just right." In the feedback condition, high/low/just right feedback appeared just above the grading sheet (Figure 18). In the control condition this space was blank.

Results: Feedback reduces grading errors

Using a repeated sampling analysis (as in Section 3), we compared staff grades to a random sampling of peer grades from participants in each condition for ground-truth submissions. The difference between the median peer grade obtained by sampling from the feedback condition and the staff-grade was 6.77%, compared to 7.74% in the no-feedback condition (Figure 13). We built a linear model that predicts grading error using experimental condition as fixed effect, and each rater as a fixed-intercept random effect.

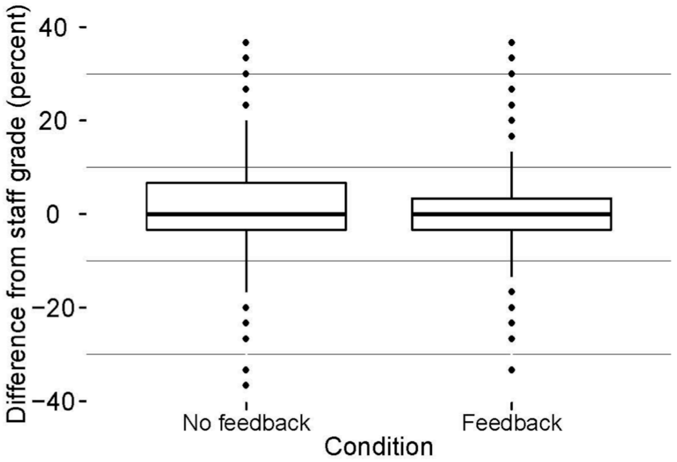
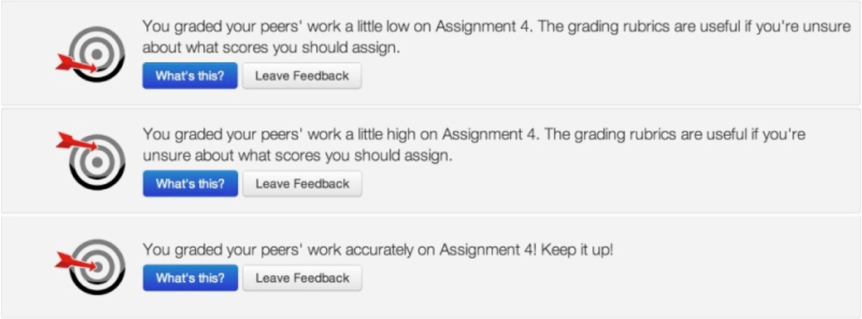


Figure 19: Feedback on grading accuracy reduced the overall error in assessment and made the range of errors smaller.



You graded your peers' work a little low on Assignment 4. The grading rubrics are useful if you're unsure about what scores you should assign.

You graded your peers' work a little high on Assignment 4. The grading rubrics are useful if you're unsure about what scores you should assign.

You graded your peers' work accurately on Assignment 4! Keep it up!

Figure 20: In the feedback condition, students received feedback about how well they were grading.

The effect of the presence of feedback is significant: t(4998) = -3.38, $p < 0.01$. 4.4% more samples in the feedback condition obtained a grade within 5% of the staff grade than those without feedback. Notably, 55 students left comments expressing their appreciation or receptiveness to this feedback; none expressed resentment.

This experiment tested the mere presence of accuracy feedback. Future work can assess the effects of richer feedback, such as the amount of bias or change over time. It can also explore bi-directional communication between the submitter and the assessor.

Providing personalized, qualitative feedback on assignments

Accurate, actionable feedback helps students improve their work [106]. Actionable feedback is most useful if it is personalized, and targets the student's recent work [107].

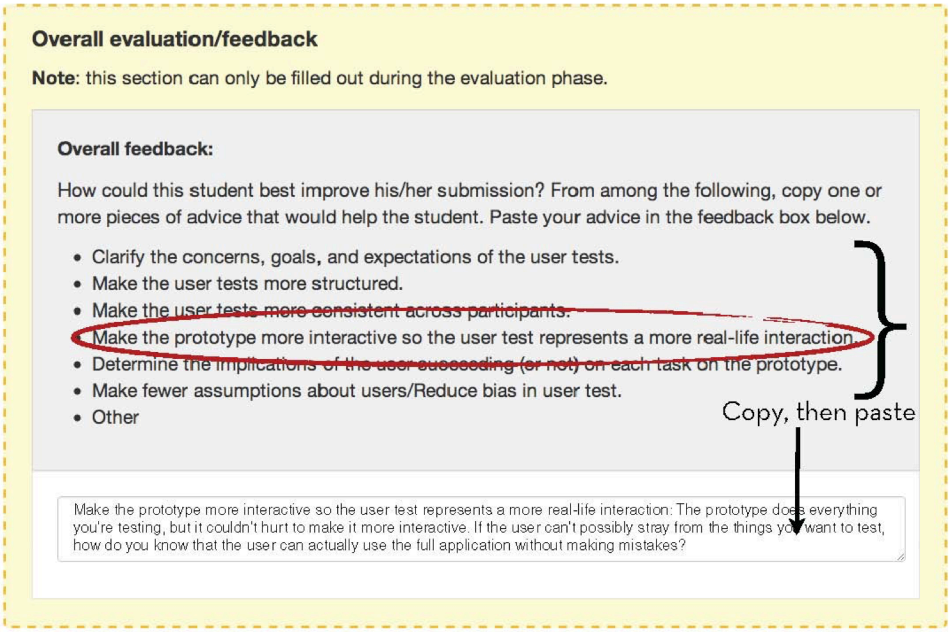
Rubrics provide feedback through quality gradations for each dimension. For instance, students can look at rubric items they did poorly on to find areas for improvement. However, using rubric item scores as feedback has two important limitations. First, students must reflect on why they did poorly on some topic. Unfortunately, these are often topics the student understood poorly in the first place. Second, rubrics only point out areas for improvement, not how to improve.

Can peers provide actionable, personalized feedback? We introduce one method that captures broadly applicable yet specific feedback in short snippets. On the assessment form, raters select which snippets apply to the current assignment, and optionally fill in a “because. . . ” prompt (Figure 21). Inspired by [108], we call the result “fortune-cookie feedback” for its brevity and general applicability. Table IV shows some examples.

Methods: Creating fortune cookies

We wanted fortune cookies to help with two common patterns in student performance. First, we wanted to find places where committed students did poorly, and retroactively generate useful advice. To find committed students (and keep the number of submissions manageable), we restricted our analysis to students whose initial performance was above the 90th percentile. Then, we compared students who subsequently got the median grade to those that got grades above the 90th percentile.

Second, we wanted to highlight strategies that students used to improve. We compared submissions from students that improved their performance from median grade to excellent (above 90th percentile) on a subsequent assignment against those that obtained



Overall evaluation/feedback

Note: this section can only be filled out during the evaluation phase.

Overall feedback:

How could this student best improve his/her submission? From among the following, copy one or more pieces of advice that would help the student. Paste your advice in the feedback box below.

- Clarify the concerns, goals, and expectations of the user tests.
- Make the user tests more structured.
- Make the user tests more consistent across participants.
- Make the prototype more interactive so the user test represents a more real-life interaction.
- Determine the implications of the user succeeding (or not) on each task on the prototype.
- Make fewer assumptions about users/Reduce bias in user test.
- Other

Make the prototype more interactive so the user test represents a more real-life interaction: The prototype does everything you're testing, but it couldn't hurt to make it more interactive. If the user can't possibly stray from the things you want to test, how do you know that the user can actually use the full application without making mistakes?

Figure 21: Students copied snippets of feedback (fortune cookies), pasted them in a textbox and optionally added an explanation.

median grades on both assignments.

We then manually wrote feedback for each submission separately. For each assignment, we looked at an average of 15 submissions; five each that showed improved, reduced and steady performance. Combining related feedback from different submissions led to our final list of warning signs and improvement strategies. Creating fortune cookies took a teaching assistant 3-4 hours per assignment.

We created fortune cookies based on submissions in the first iteration of the class, and tested them in the second iteration. As the last question on the grading sheet, we asked, “which of these suggestions would improve this submission the most?” Students copied appropriate fortune cookies from a list and pasted it in to a textbox below. Students were not required to use these snippets for feedback—they could type in their feedback into the textbox as well.

Results: How well do fortune cookies work?

Overall, 36.2% of assessments included feedback (compared to 36.4% in the previous iteration without cookies). A chi-square test on the number of assessments that contained feedback suggests that fortune cookies do not encourage more students to leave feedback (_$_2$= 0.1, p = 0.75). Because submissions were assessed by multiple students,

Table 5: Example fortune cookie feedback

Assignment	Fortune Cookie
Needfinding	Brainstorm more diverse needs
Needfinding	Brainstorm more specific needs
Needfinding	Develop a more specific point of view [for proposed solution to need]
User testing plan	Clarify the concerns, goals, and expectations of user tests
User testing plan	Make the prototype more interactive so the test represents a more life-like interaction.

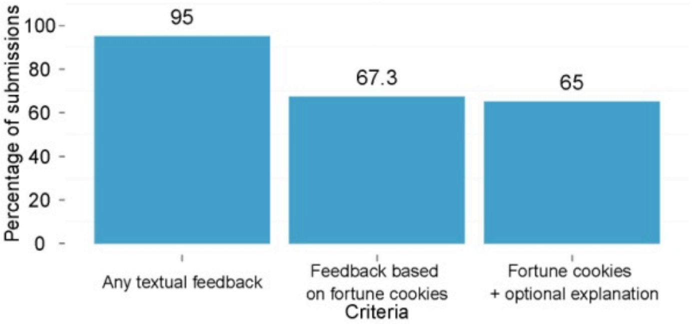


Figure 22: Most students received at least one piece of textual feedback. Most fortune cookie feedback was personalized.

94.9% of submissions received at least one piece of written feedback (compared to 83% without cookies); 67.2% of students received at least one “fortune cookie”; and 65% of students received one or more fortune-cookies with a “because. . .” explanation (Figure 22).

Raters typed the same amount of feedback whether or not an assignment contained fortune cookies. If we subtract the text of the cookie itself, there was no significant difference in comment lengths whether or not cookies were used (t(10673) = 0.44, p>0.6). If the text is included, comments that used fortune cookies were longer (t(10673) =3.61, p < 0.05). This suggests that students expend the same amount of effort writing feedback, and using fortune cookies allows this effort to be used to add to the fortune cookie text.

Discussion

Reusable pre-canned prompts encourage students to direct their effort to providing feedback beyond the cookie text. While we do not demonstrate this improves feedback, we see three reasons why fortune cookies may provide better quality feedback than non-cued feedback. First, providing raters a list of potential feedback items changes a recall/identification task into a recognition task. This reduces the cost of giving feedback [109], [110]. Second, showing a list of common, assignment-specific problems that the submission could have potentially reduces inhibition, and encourages peers to think critically [111]. Third, because fortune cookies sometimes used terminology learned in class, they may have triggered cued-recall of these concepts [112], leading to more conceptual comments.

Future research could investigate this idea further. In addition, it could also explore if fortune cookies confer differential benefits to different students and how best to leverage this.

Overall discussion

Using data to improve assessment materials

Iterative design often pays big dividends [113], and assessment systems are no exception. The large scale of online classes allows data-driven iterative improvements of classroom materials in ways that small classes may not. Below, we describe some data-driven changes we made.

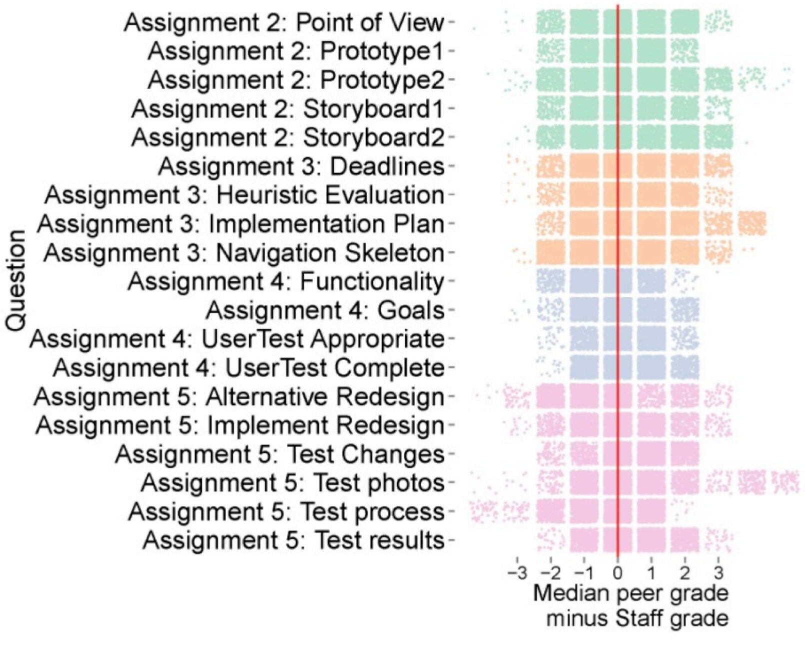


Figure 23: Comparing variance of rubric items can help teaching staff find areas that may need improvement. For example, this figure shows the variance for four assignments of the HCI course between staff grade and median peer grade. A narrow, dense band indicates higher agreement. For example, Assignment 4 (blue) has generally higher agreement.

One can use low rater agreement to find questions that might benefit from revisions. We found that peer and staff raters agreed far more on some questions than others (Figure 23), and that questions with low staff agreement also had low peer agreement (r = 0.97, t(24) = 19.9, p < 0.05). We reviewed such questions and revised them with feedback from the forum. Most rubric revisions centered on making rubrics more easily readable.

Improving readability: Some rubrics sometimes used a non-parallel grammatical structure across sentences. This is not uncommon: even examples in prior work on using rubrics suffer from this problem (e.g. [9]). We hypothesized that using a parallel sentence structure would better help students understand conceptual differences [114]. We found that rubric items with parallel sentence structure in the first iteration had lower disagreement scores (F(1,39) = 2.07, p < 0.05) (Figure 24). We revised all rubrics to use parallel sentence structure. We also made other changes to improve readability, such as removing duplicate information from assignments, and splitting up rubric items that asked students to make a complex judgment (e.g. “Is the prototype complete and functional?” to “Is the prototype complete?” and “Is the prototype func-

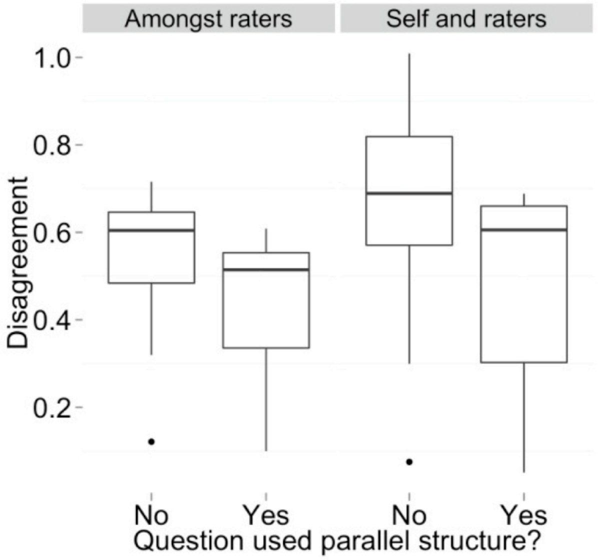


Figure 24: In iteration 1, questions with parallel structure had lesser disagreement, both amongst peer graders, and between the median grade and the self-assessed grade. We changed all assignments to use parallel structure across rubric items.

tional?”).

Word Choice: Although the rubrics had been revised for three years in the in-person class, many forum posts asked for clarifications of ambiguous words. Words like “trivial”, “interesting”, “functional”, and “shoddy” may be correctly interpreted by the on-campus student with a lot of shared context, but are ambiguous online. The revised version replaces these words with more specific ones (which may help on-campus students as well).

The revised rubrics were used in the second iteration of the class. Overall, the peer-staff agreement was 2.5% higher than the previous iteration.

Going beyond pass/fail

Peer assessment as described in this chapter works reasonably for a pass/fail class. How might peer assessment be used in classes that award more fine-grained grades? Beyond having iteratively-refined rubrics (as above), one possibility is to involve community TAs in grading submissions that are estimated to have low grading accuracy (e.g. with large differences between self and peer grades). In addition, our early experiments suggest that greater accuracy is possible by weighting different raters’ grades differently, an important topic for future work. Lastly, our experiments suggest that machine-grading approaches (such as those for essay grading) may be combined with peer assessment to provide accurate assessment. Later chapters in this dissertation provide an overview of such methods.

Inflating self-grades and other gaming

Many types of cheating are currently possible and unchecked in online classes. For example, someone else could simply take a course on your behalf. To the extent that participation in the online classroom is based on intrinsic motivations (such as a desire to learn), students rarely blatantly cheat [115]. (Anecdotally, several instructors in early online classes have reported that some students appear to be cheating, but that it doesn't currently appear to be widespread.)

To date, large-scale online classes, including our own, have primarily emphasized learning, rather than certification [65]. Students do not receive much in the way of credit. (Though on social media like Facebook and LinkedIn, some students report having "attended" Stanford.) Still, some students probably attempted to game their score by strategically over-reporting their grade (Figure 25). As online classes count for more benefits, such gaming may increase.

Gaming also has a silver lining. A valuable skill for success is the theory of mind to intuit how others perceive one's performance [28], and gaming may help students develop this skill. Cheating may also arise if the value of officially recorded performance in these classes increases (e.g. [102], [116]). To combat this, several organizations have proposed solutions like in-person testing facilities (e.g. [64]), or verified-identity certification [117]. Others remain focused on teaching for students who want to learn [65].

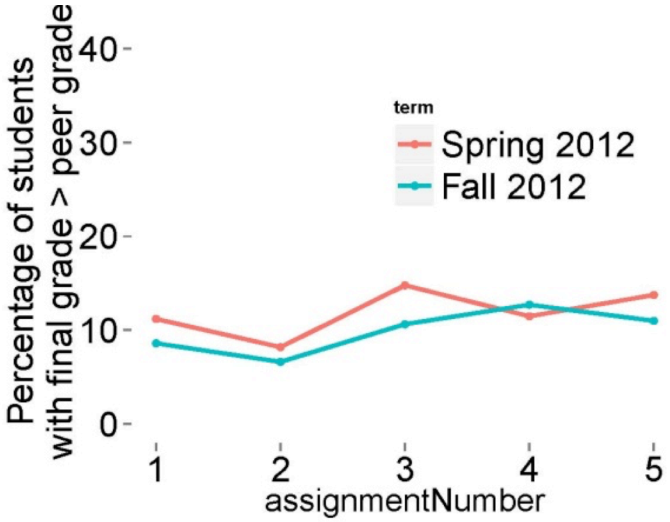


Figure 25: Students in the second (Fall 2012) iteration of the class reported a self grade > 5% higher than peer grade more frequently, and so got their self grade less frequently.

Limitations of peer assessment

While peer assessment offers several benefits, it also has limitations. First, peers and experts (e.g. staff) may interpret work differently (see Appendix A.2). Such differences are well-known in related fields: Experts and novices both robustly reach consensus about creativity, but their consensual judgments differ from each other [118]. This may be because novices and experts differ in their tacit understanding of value [119]. Peer assessment addresses this problem by providing raters with expert-made rubrics, but some differences may persist. In addition, independent assessment via rubrics and subsequent aggregation may not assess "controversial" work well.

Second, peer assessment imposes a particular schedule on class, and limits student flexibility. In our class, several students complained in class forums about being unable to complete peer assessments in time. Lastly, while peer assessment works well for the large majority of students, students who receive an unfair assessment may lose motivation. Anecdotally, we have noticed that students are generally satisfied with their overall grade, but are frustrated by inaccurate qualitative feedback from some peers. Addressing these motivational aspects remains future work.

The changing role of teachers

Peer assessment fundamentally changes the role of staff. When peer assessment provides the primary evaluative function, the staff role shifts to emphasize coaching [120]. Students sometimes believe that teachers grade on personal taste, and focus on currying favor. By contrast, when teachers coach but do not grade, students focus more on conceptual understanding [121]. Also, providing explicit grading criteria (especially in advance) helps convey to students that grading is fair, consistent, and based on the quality of their work.

Peer assessment also changes how instructors spend their time. When staff assess student work, their effort is focused on doing the grading. By contrast, with peer assessment, the instructor's main task is articulating assessment criteria for others to use.

Because of the diversity of submissions, this can be extremely difficult to do a priori. Teachers should plan on revising rubrics as they come across unexpected types of strong and weak work. After revision, these rubrics can scale well for both students and other teachers to use. For online education to blossom, it will be important to teach the teachers best practices for rubric creation, and to create effective design principles and patterns for creating assessments.

While the scale and medium of online education poses new challenges, it also offers new solutions. In key areas, online education encodes pedagogy into software, which increases consistency and supports reuse – and defaults have a powerful impact on behavior [122].

The role of teaching staff (TAs) changes too. Instead of spending a majority of their time grading, they spend a large fraction of their time fielding student questions, mentoring students, boosting student morale and autonomous perspective, and making data-driven revisions to class materials.

The changing roles of students

One of the most remarkable results from our experience was that students reported that assessing others' work was an extremely valuable learning activity. Can online classes provide an avenue not just for peer assessment, but for peer learning as well?

The second iteration introduced Community TAs recruited among students from the first iteration (Armando Fox and David Patterson's Software-as-a-Service online class used a similar program [123]). We invited students who did well in class, assessed many submissions voluntarily, and participated actively in class to become Community TAs. Community TAs volunteered their time, and were not paid. Their duties comprised grading assignments, answering student questions, and helping iteratively improve assignments. Five students from across the world participated. Together, community TAs answered 547 questions on the forum; staff (3 local TAs and the instructor) answered 582 questions. In addition to providing factual answers and assignment clarifications, Community TAs also leveraged their personal experience to offer advice and cheerleading.

We hypothesize that Community TAs are effective for the same reasons as undergraduate teaching-assistants at a university [124]. First, because community TAs had done well in the class, they possessed enough knowledge to effectively offer information and guidance. Second, because they had taken the class recently, they could easily empathize with issues students faced and also could effectively offer social support.

Massive online classes also offer individual students an opportunity to have large-scale positive impact. For example, when the first assignment of the Spring 2012 class had fewer peer assessments than needed, one student rallied her peers to finish a large number of assessments over a single day (the top ten students assessed an average of 48 submissions: nearly ten times their required number) so that students could get feedback in time. She also participated heavily in the forums, and gathered staff-like respect from her peers.

The changing classroom

The online classroom is distinctly different from its in-person counterpart. Recent research has discovered some of these differences: students in online classrooms are much more diverse both demographically, and in their objectives in taking the class, and platforms make some kinds of data, such as engagement with course material, more plentiful and finer grained, while making other information, such as facial expressions of confusion, completely inaccessible [125].

These differences require rethinking the design of the classroom. For instance, students often have work commitments, and holidays are at different times around the world. This reflects in class scheduling: the first iteration of the class spanned seven weeks, mirroring the time these topics take in the Stanford course. Although university-like deadlines helped generate interest in online classes, we found that campuspaced deadlines are too rigid online. Consequently, the second iteration spanned nine weeks to give students more time and flexibility.

While class diversity requires adaptations, it also inspires new opportunities. How can teachers support student leadership and community learning more directly in the online classroom? Again, the design studio offers inspiration [2], [11]. By making not only the results of work, but also the process of creation highly visible, it helps students learn and build awareness through observation [126]. In addition, a studio facilitates dialogue between students, instructors and artifacts that help students collaboratively learn difficult concepts and solve problems [11].

The opportunity here is two-fold. First, online learning can be blended with co-located learning. Even though this was a completely online class, students self-organized to meet up in ten locations around the world including London, San Francisco, New York City, Buenos Aires, Aachen (Germany), and Dhaka (Bangladesh).

Second, we can build online experiences that are inspired by the physical studio. By removing the constraints of the physical classroom, online classes have made education accessible to many new kinds of students—the new mother, the full-time professional, and the retiree. Preserving this accessibility, while providing the benefits of the in-person classroom online offer a promising area for future work.

More generally, online education requires us to re-conceptualize what it means to be a student in many ways. One has to do with enrollment and retention [35]. Typing one's email address into a webpage is not the same as showing up for the first day of a registrar-enrolled class. It's more like peeking through the window, and what the large number of signups tells us is that lots of people are curious. How can we convert this curiosity into meaningful learning opportunities for more students?

Conclusions and future work

This chapter described our experiences with the largest use of peer assessment to date. This chapter also introduced the “fortune cookie” method for peers to provide each other with qualitative, personalized feedback. We demonstrated that providing students feedback about their rating bias improves subsequent accuracy. There are many exciting opportunities for future work.

First, systems could allocate raters and aggregate their results more intelligently to increase accuracy and decrease work. Crowdsourcing techniques suggest initial steps. After assessment is complete, systems could differentially weight grades based on raters’ past performance, for instance, extending approaches like [127]. Also, the number of raters could be dynamically assigned to be the minimum required for consensus, extending e.g. [128]. Furthermore, an algorithm could adaptively select particular raters based on estimated quality, focusing high quality work where it’s most needed, as in [129]. Finally, as with standardized essay grading [67], peers could be used together with automated grading algorithms (such as [130], [131]). This hybrid approach can achieve consensus while minimizing duplicated effort. Ideally, these grading schemes should be understandable as well as accurate. Should the system show students how their grade was generated? And if so, how?

Second, current online learning platforms suffer from sensory deprivation relative to a human teacher. They receive final work products, but have no knowledge of students’ process. Cognitive tutoring software has shown that attending to students’ process can improve learning through personalization—adapting questions, pacing, and guidance [132]. Integrating rich learner models with peer assessment offers many exciting opportunities.

Third, physical universities employ many structural levers to keep students motivated and engaged. In our experience, only a quarter of approximately 3000 students who completed a time-intensive first assignment did all five assignments. Needless to say, at a physical university the completion rate for an equivalent class is much higher.

How can online settings provide greater motivation support? Future work could draw both on research on commitment strategies in online communities (e.g. [133]) and resources used at physical universities, such as mentoring and orientation courses [134]. More generally, online learning platforms could benefit students by incorporating known best practices about learning and moving to a more evidence-based approach.

Fourth, peers can help instruction itself. One promising approach is to use social mechanisms to highlight good student work and build connections, such as [135]. Another is to leverage peers in physical meet-ups to augment instructor teaching [136]. This approach also creates technology and pedagogy design opportunities for a "flipped" classroom—what should class time look like at a university when students can watch the professor on video? Already, several universities are teaching physical classes augmented with online materials [137]. How would different roles change with such a model?

Fifth, future work has the potential to tie student work in class to skilled crowd work [138]. For instance, students in the HCI class could build prototypes and design websites for clients, or students studying Machine Learning could compete to build predictive models. How can the pedagogical goals of the class be intertwined with potentially productive work? This future work will offer students around the world an opportunity to learn in ways previously impossible.

Chapter 4 Automated systems reduce busy-work in peer assessment[6]

Short answer questions: flexible and pedagogically meaningful, but time-consuming to assess

Short answer questions are a powerful assessment mechanism. Many real-world problems are open-ended and require students to generate and communicate their response. Consequently, short-answer questions can target learning goals more effectively than multiple choice; instructors find them easier to construct; and short answers are relatively immune to test-taking shortcuts like eliminating improbable answers [139].

Many online classes could adopt short-answer questions, especially when their in-person counterparts already use them. However, staff grading of textual answers simply doesn't scale to massive classes. In our experience, grading each answer takes approximately a minute. Grading a hundred students is feasible, taking two hours per question. For an online class of 5,000 students this involves two person-weeks of grading per question. Automated grading and peer assessment both offer ways to scale assessment [37], [140], but in isolation, both introduce an unsatisfactory tradeoff.

While algorithmic grading consistently applies criteria to all student work [140], it has many shortcomings. It frequently relies on textual features [141], rather than semantic understanding. For instance, automated essay scoring software uses counts of bigrams and trigrams (sequences of two or three words) [142]; NLP techniques like syntactic parsing [143]; dimension reduction techniques such as PCA [144]; or a combination of

[6] A version of this chapter was originally published as an article in the proceeding of the ACM Conference on Learning at Scale, 2014 as [172].

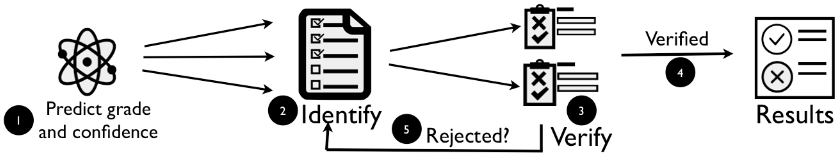


Figure 26: Overview of the assessment process. (1) Machine learning algorithm predicts grades and confidence. Number of independent identifications decided based on confidence (2) Peers identify attributes in answer using rubric (3) Two other peers verify existence of attributes. Final score is sum of verified attributes (5) if attributes are rejected, one more rater is asked to Identify. If two independent identifications are identical amongst raters, one is considered a verification (4).

these features [145]. This reliance on textual features reflects algorithms' limited ability to capture the semantic meaning of student work. This limited understanding can cause grading errors because answers using unconventional phrasing may be penalized. Furthermore, students may game algorithms with answers that match patterns, but are otherwise incorrect [146] This has, in turn, led to public skepticism about algorithmic grading [147].

Algorithmic grading for short answers is especially challenging, because the limited text provides fewer lexical features. Algorithms can still use features like word overlap, but accuracy suffers [148].

In contrast, peers can more robustly handle ambiguity and differences in phrasing, and students learn by assessing others' work. However, peer assessment requires students to spend time grading several (e.g., five) peers. Student raters need training, and still may differ in how they apply grading criteria, and ratings may drift over time [140]. Raters also suffer from systematic cognitive biases including the Halo Effect (wrongly generalizing opinions on one characteristic to the entire answer), stereotyping (e.g. gendered/nationalistic cues affect grading[37]), or perception differences (grading of prior answers affects grading of the current answer[140].

Could machine-learning algorithms mitigate grader biases and minimize human effort? Crowdsourcing algorithms can correct inter-rater differences[14], and recruit more raters when they encounter unreliable raters[149], [150]. Inspired by these successes, this chapter introduces a workflow that intelligently combines algorithmic and peer assessment to provide the benefits of both, while mitigating their individual drawbacks.

The **identify-verify** workflow uses algorithmic grading to estimate how many independent peer assessments are needed. The algorithm estimates “ambiguity” of the answer using its prediction confidence. More raters are assigned to highly ambiguous answers and fewer to less ambiguous ones. In this chapter, the range was 1 to 3 raters. Peers then identify key features of the answer using a staff-provided rubric. Other peers verify whether these feature labels were accurate. Few peers are needed when initial human ratings agree with a high-confidence machine rating. The algorithm seeks more assessments when raters disagree. The algorithm automatically seeks higher quality assessment if more raters are available.

An experiment compared hybrid grading with peer grading; 1370 students from an online human-computer interaction class participated. Compared to a baseline of aggregating independent peer ratings using a median, integrating machine grading yields comparable accuracy with lower effort. For binary questions, using the machine grading with identify (and no verify step) yields 83% of the peer-median accuracy, and only needs 54% of human effort. For an enumerative short-answer question, 70% of the effort yields 80% accuracy. For both types, adding verification yields higher accuracy and more reliable information about the answers' attributes, but increases human effort. A follow-up experiment investigated how identify-verify works with a varying number of graders, compared to the baseline of median of peer grades. Adding the verify step yielded a 20% gain in accuracy over the peer-median method with four raters.

In addition to saving time, this hybrid also provides students richer, structured feedback about their answers in addition to their scores. Students see both a list of features of the answer they got right, and common errors they made.

This chapter makes two contributions. First, it introduces the identify-verify pattern for combining peer and machine grading. Second, it presents experimental results demonstrating the accuracy benefits and the tradeoffs in human effort of the identify-verify pattern in various configurations.

Class setup

We evaluated the identify-verify approach in a large, online class introducing human-computer interaction. This class is based on an in-person class that uses short-answer questions to assess students' knowledge. For instance, short answers assess if students can construct well-formed interview questions, if they understand prototyping strategies, and can explain differences between experimental designs. The system described in this chapter introduced these short-answer questions to the online class. Students answer short answer questions on two quizzes, one in Week 3 of the class, and once on the final (Week 9).

Pilot: lenient peers, strict machines

We piloted short-answer questions in the May 2013 offering of the class. The pilot explored whether simply combining peer and machine scores using a median yielded accurate results. In addition, it aimed to understand the relative merits of machine and peer grading.

Three independent peer raters scored each student answer. The site provided raters with a grading rubric and staff-graded examples to calibrate themselves (similar to Calibrated Peer Review [94]). After grading a staff-provided example, students assessed peer answers. A machine classifier reliant on textual features scored all answers as well. The system combined human and machine scores by taking the median of all four scores. Other methods of combining grades, such as linear regression, were sensitive to outliers.

To assess accuracy, we compared the median grade to the staff grade for 200 submissions. We found that accuracy increased with increasing number of peer raters, consistent with prior work [37], [151]. In addition, we made the following observations:

- **Peers were more lenient than staff, and writing fluency swayed judgments on correctness**: Peers sometimes awarded points to plausible-sounding but incorrect answers. For instance: “Rewrite the interview question `Do you like the WordArt feature from Microsoft Word?' to address problems with it”. The problems with the interview question are that it is leading and it assumes users have an opinion on the feature. One incorrect student answer was “With respect to your experience, how much do you like the WordArt feature, on a scale of 1-5?” Three peer raters marked this as correct, even though it has the same problems as the original question. We also found that cues such as how confidently the answer was written, or whether it used fluent language seemed to affect the peer's rating. Prior work has shown similar Halo effects influence human “grading more generally [140].

- **Peers understand ambiguous answers better**: For example, for the same WordArt question, machine grading marked the correct answer “How do you add images or text in different styles into your documents in Microsoft Office?” as incorrect (possibly because training examples had few correct answers without the word WordArt). However, two of three peer raters marked it to be correct.

Together, these two factors meant algorithmic grading was stricter, since it only awarded credit when the answer matched example answers closely (the average machine grade was 16% lower than staff). Peer grading was more lenient than staff: the average peer grade was 14% higher than staff.

- **High-confidence predictions from machine grading were generally accurate, and agreed with peer assessment.** For binary questions, when the algorithm reported confidence larger than 80%, staff and machine grades matched 85% of the time (staff and a single peer agreed 78% of the time). In addition, for low-confidence predictions, staff/machine disagreement was larger than staff/median-peer disagreement. (When confidence was 50-60%, staff and machine grades agreed 53% of the time. For these same submissions, a single peer agreed with staff grade 52% of the time, but the median of three raters agreed with staff 68% of the time.) Therefore, low-confidence predictions are somewhat informative, but cannot be trusted reliably.

This pilot suggests that few peers are needed for answers graded with high algorithmic confidence, but more peers may be necessary for assessing questions with low confidence. However, a simple median for combining human grades and machine grades cannot handle machine grades is not uniformly reliable. This suggests that a grade-combination scheme should tune the number of raters based on algorithmic confidence. Essay scoring on standardized tests uses one such scheme: the GMAT compares a human essay score with the machine score, and recruits more human raters if the scores differ [152].

Combination schemes could also leverage peers' ability to understand ambiguous answers, but should account for them being biased and lenient. Prior work suggests it is possible to create processes that mitigate cognitive biases [153], [154], but simply alerting students to their biases does not help mitigate them [155]. Therefore, this chapter seeks to create a workflow and interface to mitigate biases and improve accuracy.

The Identify/Verify architecture

Based on these pilot insights, we designed a grading system to combine the strengths of human and machine grading. This system seeks to minimize human effort while still retaining current accuracy. We choose to reduce human effort, rather than improve accuracy, because many large, online classes (including our evaluation class) are pass-fail, and we found accuracy from the pilot (between 67% and 82%) reasonable. At this accuracy, we estimate the number of students who should have passed but didn't due to grading errors to be less than 3%. This chapter leverages the insight that partitioning tasks so people can audit each other improves quality and efficiency [156], [157].

Identify-verify comprises three steps (Figure 26). First, a machine-learning algorithm predicts a grade and confidence score for each submission. The system assigns a number of peers to grade the answer based on the confidence score. Second, peers use a grading rubric to *identify* which features the answer contains (Figure 27). Third, they *verify* other peers' feature identification for other answers (Figure 28). Identify-verify assigns a final grade by combining the grade for verified features in the answer; our prototype uses the sum of feature grades. For instance, if a student submission is identified to have two features each worth one point, the submission is awarded two points, the sum of feature scores. Below, we describe each step in the assessment process.

Step 1: Algorithm estimates grade and number of raters

Before peer assessment begins, a machine-learning algorithm predicts the grade for each answer. We built a generic text classifier using etcml.com with the predicted grade as the output. This classifier uses textual features such as word, bigram and trigram counts, length of answers, and letter n-grams (to capture use of word fragments like “creati-”, which match “creativity”, “creative”, “creation” etc.).

Teaching assistants provided numeric scores and correct/incorrect attributes for about 500 student responses per question. The numeric grades were used as labels to train the classifier. Instructors provided teaching assistants an initial rubric for grading. TAs then expanded this rubric with correct/incorrect attributes they identified, and added example student answers with those attributes. Future work could bootstrap attributes and examples using prominent features from the trained classifier.

The system then uses the classifier trained on staff-graded answers to grade all answers. The classifier outputs the most likely grade (the prediction), as well as the probabilities of all possible grades (e.g., an answer may have a grade of 1 with probability of 0.2, and a grade of 0 with probability 0.8). For the rest of the grading process, we use the probability of the most likely grade (in our example 0.8) as the algorithm's confidence in the grade. (Future work could consider using other statistics).

The algorithm's confidence determines the initial number of peer raters assigned to each answer. The intuition behind this is that confidence represents a measure of ambiguity---answers with high confidence are usually those that are clearly right or wrong. Conversely, ambiguous answers often have low confidence, and therefore should have more independent human assessments. We require answers with high confidence (>90%) to have a single rater, those with medium confidence (75%-90%) required two, and all other answers required three raters. Overall, 34% of student submissions had grades predicted with $>80%$ confidence, and 16% of submissions had grades predicted with $>90%$ confidence.

This chapter seeks to demonstrate the feasibility of combining human and machine grading. It does not determine the most suited machine-grading algorithm. Therefore, while our classifier represents the state-of-the-art in text classification, it does not use any special logic for answer grading.
We hope that demonstrating feasibility with a generic classifier will also inspire other researchers to create better ones.

Step 2: Peers identify answer attributes

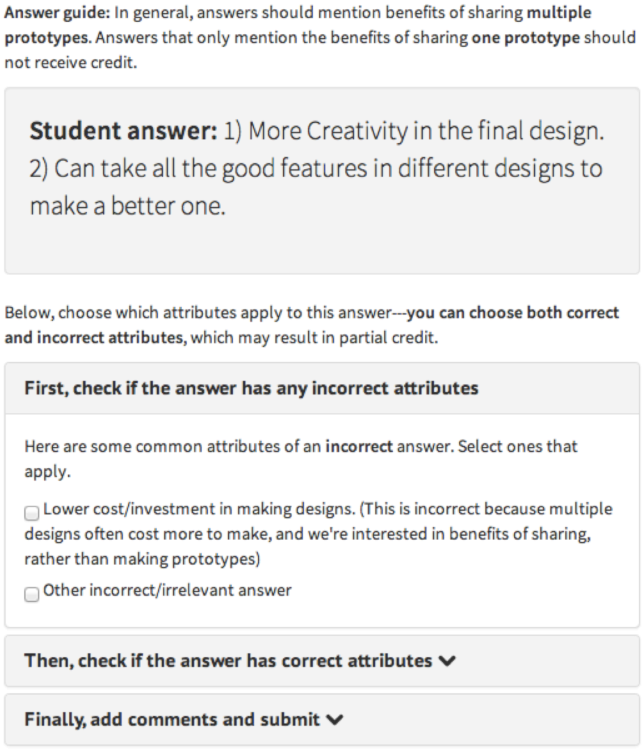
In this step, randomly chosen peers independently identify correct/incorrect attributes in student answers. Raters select these attributes from the expanded grading rubric from Step 1 (Figure 27). Staff associated a score with the presence of each attribute, which could be negative.

To minimize the impact of too-few ratings, the system solicits ratings in order of greatest need. Specifically, the system finds the student answer that has the largest number of required assessments, with the fewest completed. Ties are broken randomly.

The grading page displays this answer along with the grading rubric. Peer raters mark each attribute present by clicking a checkbox next to it. To encourage students to be critical (and reduce the leniency we saw in our pilot), the grading rubric is initially shown with incorrect attributes displayed, and correct attributes collapsed (Figure 27). Raters expand the correct attribute section by clicking the drop-down arrow.

Raters are asked to identify attributes in four student submissions. After a rater completes identification, the answer and its attributes are queued for verification. If two identifiers independently select the same attribute, that also constitutes verification. Such answers skip the separate verify step.

Even with high-confidence machine predictions, it is important that student grades do not suffer due to an over-optimistic algorithm. The current system requests one addi-



Answer guide: In general, answers should mention benefits of sharing **multiple prototypes.** Answers that only mention the benefits of sharing **one prototype** should not receive credit.

Student answer: 1) More Creativity in the final design. 2) Can take all the good features in different designs to make a better one.

Below, choose which attributes apply to this answer---**you can choose both correct and incorrect attributes,** which may result in partial credit.

First, check if the answer has any incorrect attributes

Here are some common attributes of an **incorrect** answer. Select ones that apply.

□ Lower cost/investment in making designs. (This is incorrect because multiple designs often cost more to make, and we're interested in benefits of sharing, rather than making prototypes)

□ Other incorrect/irrelevant answer

Then, check if the answer has correct attributes

Finally, add comments and submit

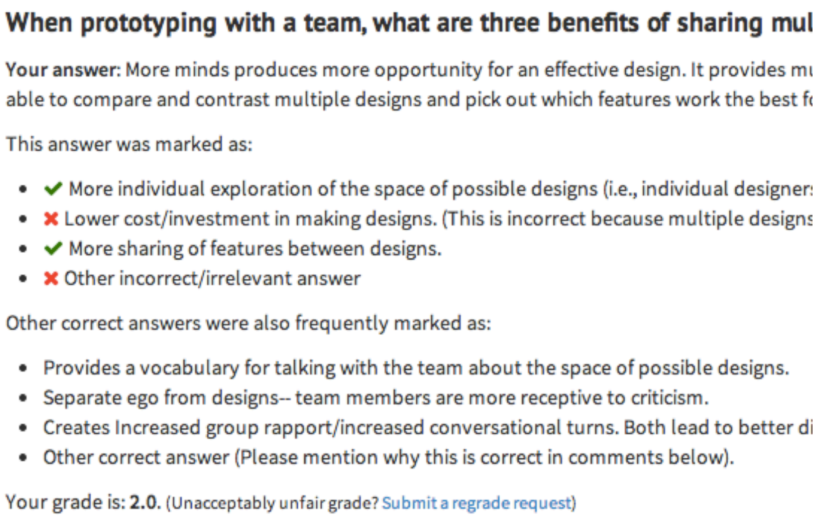
Figure 27: Identify UI: Students identified whether student answers had staff-provided features (which indicated right/wrong answers.)

tional identification for high-confidence answers where the peer and algorithm grades differ by one or more points. (In this chapter, answers are worth up to 3 points, and only whole point values are awarded.)

Step 3: Other peers verify attributes correctly identified

Now, independent raters verify attributes identified in the previous step by other peers. This interface groups answers according to the identified attribute, e.g. grouping all answers marked as “More sharing of features between designs” (Figure 28). Peers then verify whether answers contain the marked attribute. We hypothesize that grouping submission marked with the same attribute increases accuracy because verifiers are presented with a group of nominally similar responses for comparison.

When two raters independently verify an identified attribute, the system marks the attribute as verified and removes it from the verification pool. If two raters reject an identified attribute, the system returns the submission to the identify pool for one additional identifier, since the initial identification was inaccurate.



When prototyping with a team, what are three benefits of sharing mul

Your answer: More minds produces more opportunity for an effective design. It provides mu able to compare and contrast multiple designs and pick out which features work the best fo

This answer was marked as:

- ✓ More individual exploration of the space of possible designs (i.e., individual designer
- ✗ Lower cost/investment in making designs. (This is incorrect because multiple designs
- ✓ More sharing of features between designs.
- ✗ Other incorrect/irrelevant answer

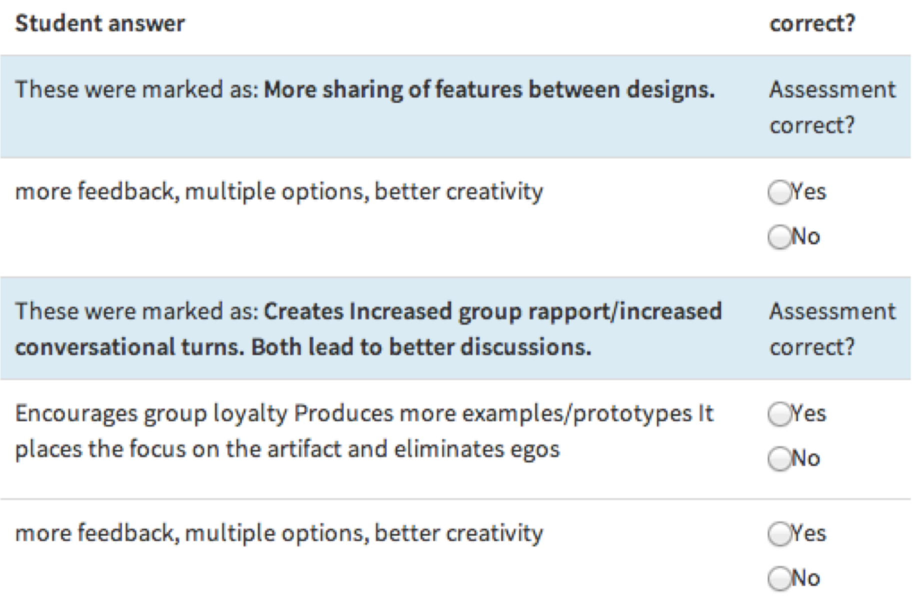
Other correct answers were also frequently marked as:

- Provides a vocabulary for talking with the team about the space of possible designs.
- Separate ego from designs-- team members are more receptive to criticism.
- Creates Increased group rapport/increased conversational turns. Both lead to better di
- Other correct answer (Please mention why this is correct in comments below).

Your grade is: **2.0.** (Unacceptably unfair grade? Submit a regrade request)

Figure 29: Identify-verify presents student grades with features present, and those missing in answers.

Similar to the identification step, the system presents submissions to verifiers in decreasing order of the number completed, and breaks ties randomly. This again provides every submission with some data quickly. This algorithm also needs at most three verifications: after three, each attribute will either have been verified, or rejected.



Student answer	correct?
These were marked as: **More sharing of features between designs.**	Assessment correct?
more feedback, multiple options, better creativity	○Yes ○No
These were marked as: **Creates Increased group rapport/increased conversational turns. Both lead to better discussions.**	Assessment correct?
Encourages group loyalty Produces more examples/prototypes It places the focus on the artifact and eliminates egos	○Yes ○No
more feedback, multiple options, better creativity	○Yes ○No

Figure 28: Verify UI: Students verified if other peers had assessed answers correctly.

Optimizing the number of raters

Identify-verify reduces the grading workload by recruiting fewer raters when the grading algorithm reports high confidence. This scheme is also cautious. First, we increment the number of identifications required for high-confidence predictions if peers disagree with the predicted grade. Second, identified attributes for an answer that are rejected may indicate the answer was difficult to grade, so we request additional assessments.

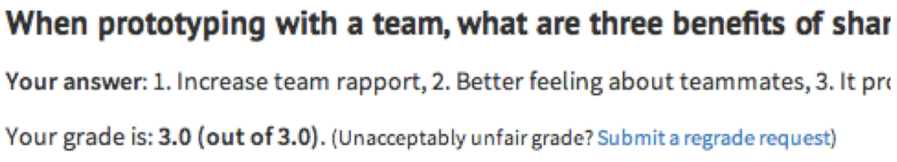
Display results and feedback

A student's final score is the sum of scores of all verified attributes, clamped to the minimum and maximum score for the question. Students see their score along with the features that peers identified, and correct attributes that their answer missed (Figure 29). Thus, students receive more than a grade: they receive detailed information about what they did well and poorly.

Evaluation

Identify-verify seeks comparable accuracy to using the median grade of independent peers, but with less human effort. Our comparison baseline asks three peers to grade a student answer.

Experiment 1: Does identify-verify yield accurate



When prototyping with a team, what are three benefits of shar

Your answer: 1. Increase team rapport, 2. Better feeling about teammates, 3. It pr

Your grade is: **3.0 (out of 3.0)**. (Unacceptably unfair grade? Submit a regrade request)

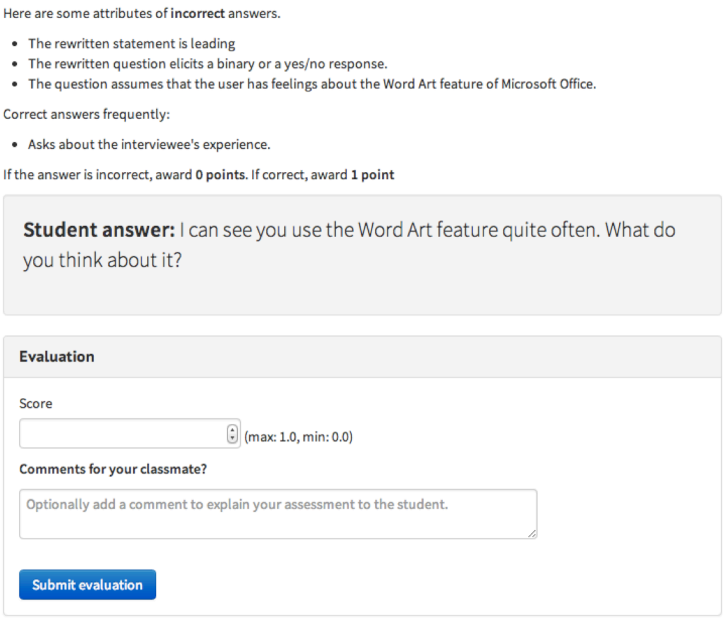
Figure 30: Student grade display in baseline condition (Grades are computed using Identify-verify, but detailed feedback is hidden.)

grades?

This controlled experiment explored two questions: First, does identify-verify grade accurately and lower effort? Second, does identify-verify reduce leniency from our pilot? (We hypothesize that leniency is due to the Halo effect, and using a structured process and interface would reduce this bias [154].)

Conditions

This between-subjects experiment had three conditions. In the *peer-median* condition, students assess four peers using a grading rubric, and enter their grade into a text field (Figure 31). In the *identify-only* condition, students assess four peers using the same grading rubric, but would instead use the Identify interface to select which aspects of the rubric were present in the student answer (Figure 27). In the *identify-verify* condition, students assessed four peers using the Identify interface. Then, they would verify assessments of eight answers that other students had created in the Identify step (Figure 28).



Here are some attributes of **incorrect** answers.

- The rewritten statement is leading
- The rewritten question elicits a binary or a yes/no response.
- The question assumes that the user has feelings about the Word Art feature of Microsoft Office.

Correct answers frequently:

- Asks about the interviewee's experience.

If the answer is incorrect, award **0 points**. If correct, award **1 point**

Student answer: I can see you use the Word Art feature quite often. What do you think about it?

Evaluation

Score

(max: 1.0, min: 0.0)

Comments for your classmate?

Optionally add a comment to explain your assessment to the student.

Figure 31: Peer-median UI: Students entered grades in a text box.

We wanted to reduce grading burden in the class, and since we hypothesize that Identify-verify would save student effort, the experiment used an unbalanced assignment; 20% of students randomly assigned to the *peer-median* condition, and the rest split evenly between *identify* and *identify-verify*.

Questions

Students assessed answers to two short-answer questions. Question 1 asked students to rewrite an interview question: "Rewrite the following interview question to address its problems: 'Do you like the Word Art feature of Microsoft Office?'" and had a binary grade (credit or no-credit). Question 2 asked students to enumerate "three benefits of sharing multiple designs with your team members, instead of sharing only one design?" Students could earn 0-3 points on this question, one per enumerated benefit. Students assessed four submissions per question, so there were a total of eight assessments per participant.

After they had completed grading, the system invited students to participate in a short survey. The survey measured trust in the system, and time taken for grading vis-a-vis their initial expectations.

The system showed students their final grades a day after the peer assessment period ended. All students saw grades computed using Identify-verify. To measure the effects of detailed feedback, the system showed those in the peer-median condition only the final score (Figure 30), students in other conditions saw both the score and identified attributes (Figure 29). After they saw results, we invited students to a second survey, which gauged how accurate they perceived grading to be and how satisfied they were with feedback.

Participants 2,556 students submitted answers; 1,370 performed assessment (the others dropped the class). 620 students participated in the pre-results survey, and 102 participated in the post-results survey. In all, students created 11006 assessments and 12264 verifications.

Measures

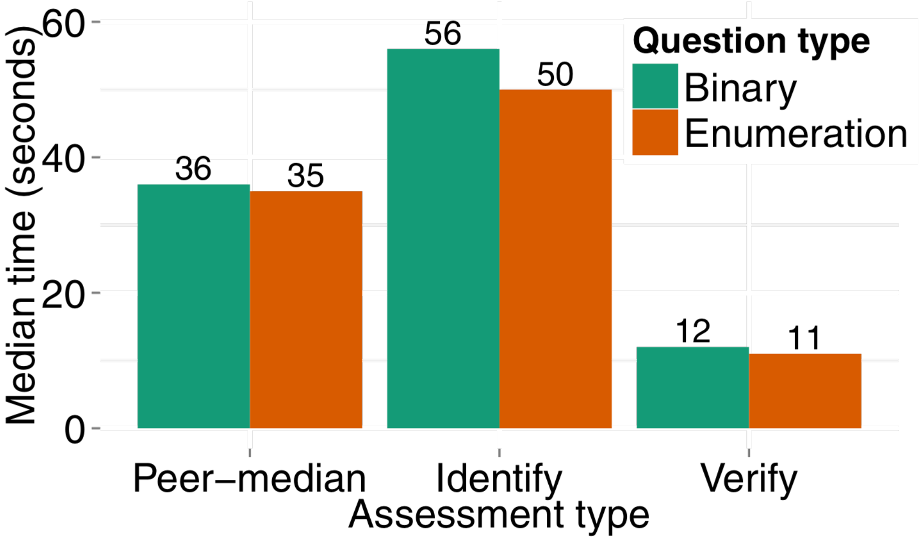


Figure 32: Assessment took longer using the Identify interface, but yielded more accurate results.

For both the *peer-median* and the *identify-verify* strategies, course staff looked at 100 student answers for each question with three peer-median assessments, and 100 more answers with two peer-median assessments. (We did not select based on the number of identify assessments, because the system dynamically determined this number for each answer). For each student answer, we compared the staff grade to the computed grade.

Results

In terms of both effort and accuracy, the ranking of conditions was the same: *Peer-median* was highest; *identify-verify* was the middle, and *identify-only* least (See Figure 32.) *Peer-median* had three raters. *Identify-only* had median one rater. *Identify-verify* had median one rater, with two verifiers for the binary question and three verifiers for the enumeration.

How accurate is identify-verify assessment?

Peer-median required disproportionately more effort than *identify-only* to achieve its results. *Identify-only* consumed 54% of the effort to achieve 83% of the accuracy in the binary question, and 71% of effort for 80% of accuracy in the enumeration question. Identify-verify consumed 84% of effort for 85% of accuracy in the binary question, and identical effort for 92% of accuracy for the enumeration question. This study only examined one effort level. The second study simulates multiple effort levels.

Verification provided a large benefit for the enumeration question, but minimal benefit for the 1-level question. Labels were rejected at similar rates (19.8% for 1-level and 18.6% for enumeration). For a binary question, not all attributes need to be identified to accurately grade it (for example, if the answer is wrong for two reasons, identifying just one is sufficient). Therefore, we hypothesize that the benefits of verification are larger for questions that are non-binary, and investigate this in Experiment 2.

Identify assessments take longer, more accurate

Students took significantly longer to select an attribute label than to select a score (see Figure 32), log-transformed t(6789)=28, p<0.01). Labeling also yielded more accurate work (see Figure 32). Identify-verify reduced leniency, while retaining peers' ability to assess unusual answers better than machines (see Table 6 and Table 7).

Identify-verify reduces voluntary acceptance

Fewer students in the *identify-verify* condition reported wanting to continue using the grading interface for other quizzes (64% said yes, t(732)=2.9,p<0.01); no significant differences existed between *peer-median* and *identify-only* (78% and 75% respectively). Usability challenges with the verify interface may have reduced interest. Some students reported that the “the layout was very confusing” others were initially unsure if they were verifying the student answer or the label. 15.8% of students in the *peer-median* condition completed more assessments than required, while 8% of students in

Table 6: Peer grade averages in points. Identify-verify reduces leniency compared with peer-median.

Question	Peer-median 3 raters	Identify-verify	Staff	Machine
Yes/no (1 point)	0.57	0.33	0.31	0.17
Enumeration (3 points)	2.17	1.65	1.74	1.35

the *identify-only* condition completed more than required.

Fewer students in the *identify-verify* condition believed the process would give them a fair grade (Asked as Yes/No: **β**=0.12, t(734)=2.7,p<0.05). This may be because verify explicitly revealed individual peers work; reducing trust. One student said that based "on the verification step of the peer assessment I'm not confident that people's quizzes are being assessed correctly." Furthermore, *identify-only* students reported more accurate grades ($\mu=1.9, t(93)=2.04, p<0.05) than those in the *peer-median* or *identify-verify* conditions (\mu=2.5, 4-point Likert scale with 1: "very accurate").

Experiment 2: How number of raters affects accuracy

A second experiment investigated how the number of raters affects accuracy. As before, students were assigned to either the *identify-verify*, *identify-only* or the *peer-median* condition. All raters graded one of fifty randomly selected submissions. 634 students participated.

The final had three enumeration questions asking students to a) mention one disadvantage of a between-subjects experimental design, b) list three ways of visually grouping related information, c) list two situations where heuristic evaluation is preferable to user testing. The experimental setup was identical to Experiment 1.

Measures

We performed a bootstrapped simulation of the peer assessment. This simulation chooses a random sample of raters for each question. We then calculate the final grade using ratings only from this sample of raters, and compare it with the staff-assigned grade. Repeating this process multiple times estimates peer agreement with staff [37]. Figure 33 shows median results from 20-repetition sampling, with one to eight raters.

We benchmark each condition against its peak accuracy: the highest accuracy seen in that condition in our simulation. More raters did not always improve accuracy, so peak accuracy was achieved with fewer than eight raters in the *identify-only* and *peer-median* conditions.

Results

A few raters identify most features

A small number of raters can identify most attributes present. Figure 33 shows that accuracy quickly plateaus, and four raters yield 92% of the peak accuracy with the *identify-only* method. Overall, the peak *identify-only* accuracy was 55% with six raters; the *peer-median* had a peak accuracy of 66% with seven raters. This early saturation is similar to heuristic evaluation of interfaces [113], suggesting similar processes may be involved.

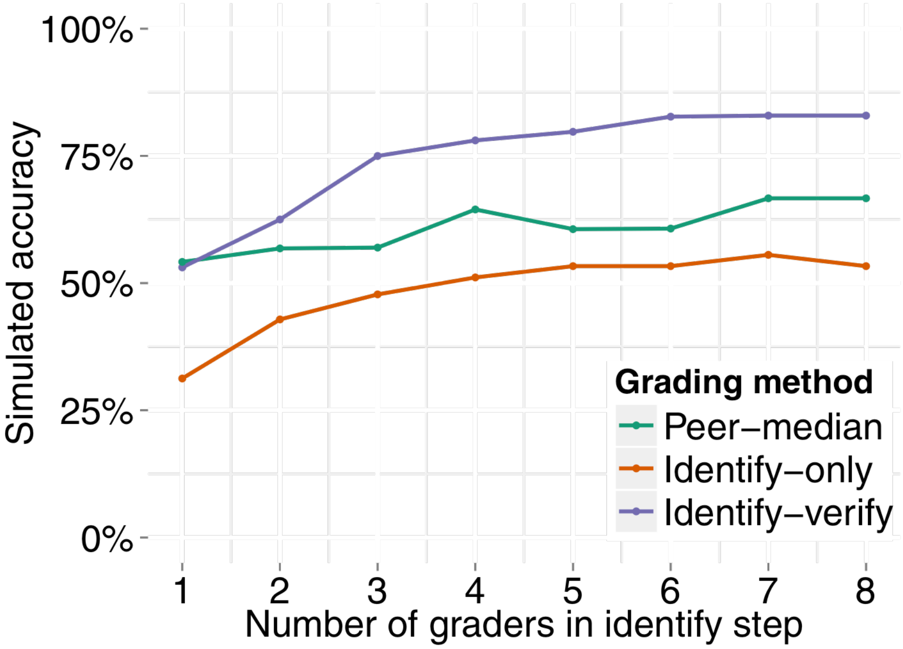


Figure 33: For enumeration questions, identify accuracy is lower than the peer-median method. Identify-verify obtains better accuracy than peer-median, especially with three or more raters.

Identify raters satisfice, Identify-only errors accumulate

Identify-Only accuracy was lower than *peer-median*, and much lower than *Identify-Verify* (see Figure 33).

First, most raters select only one attribute, even though the answer may match multiple attributes. Of the 1488 assessments collected, only 173 had more than one selected attribute. In contrast, staff assessments averaged 1.4 selected attributes. Second, because identifiers sometimes mislabel answers and there is no mechanism (i.e. verification) that catches this, asymptotically optimal performance is with relatively few raters and relatively low quality. In contrast, the peer-median approach uses the median of peer grades in the peer-median approach, so grades become more accurate with more raters as outlier ratings are discarded.

Many identifiers appear to have selected the first relevant label (Figure 34). Random-

Table 7: Sampling of errors in assessment. Peer ratings help when machines are less confident of the grade.

Student answer	**Remarks**
"How do you use the Word Art feature and how does it help you to meet your goals?"	Machine marked as incorrect, possibly because of leading bigrams "does it", "help you". Peers marked as correct. Staff graded as correct.
"What do you think of the Word Art feature of Microsoft Office?"	Construction marked as incorrect in the grading rubric (because it assumes opinion); yet, two of three peers in the peer-median condition marked as correct (possibly because it's less leading than "do you like…"). Both machine, and identify peers marked as incor-rect.
"What would you like to see changed in the `Word Art' feature on Microsoft Of-fice?"	Possibly useful interview question asks how to change, instead of under-standing current use (and so, is wrong): 3 peers in the peer-median condition marked correct; one rater identified it as `Other correct answer', but verification rejected it. Staff graded as incorrect.
"Inspiration. Innovation. Social" (for benefits of sharing prototypes)	Uses keywords without context. Machine awarded one point (possibly due to 'Inspiration'), but Identify peers did not (this answer had no peer-median assessments), nor did staff
"Because the best way to have a good idea is to have lots of ideas." (For benefits of sharing prototypes)	Pithy and plausible, but irrelevant. Awarded 1 point (out of 3) in peer-median evaluation, none in Identify. Staff graded at 0.

izing order across raters should mitigate ordering effects. Future work could investigate interfaces that incent raters to select all relevant labels.

Verification improves accuracy, especially with more raters

Identify-verify yielded the highest accuracy: the peak accuracy was 82% with six raters. The simulation required labels to have one peer verification and no peer rejections. (Actual student grading requires two verifications. Because the system solicits verifications in decreasing order of need, the median staff-graded submission had only one verification, or was rejected.)

Even single-peer verification dramatically increases accuracy. With three raters, accuracy is 28% higher than *identify-only*, and 18% higher than *peer-median*. *Peer-median* assessments took a median time of 19 second, identifications took 40s. Verification took 12s, similar to Experiment 1. Therefore, this 18% boost in accuracy comes with approximately two extra minutes of human effort per answer.

Because verification filters out erroneous identifications, its benefit is larger with more raters: verification with one rater yields a 22% benefit in accuracy, with four raters, it yields a 27% benefit. In our simulation, three identifiers identified most attributes, and inaccuracies with three or more raters are due to wrongly identified attributes.

Discussion

Identify-verify represents one choice in the trade-offs between human effort and grading accuracy. This choice was optimized for a large, pass-fail class.

Is verification necessary?

Our results demonstrate how erroneous identification can be detected with an easier operation (verification), similar to Soylent [156]. This is especially useful for questions where all attributes need to be correctly identified. While verification increases grading time, it yields more yields more descriptive, actionable, and accurate student feedback, which helps students learn.

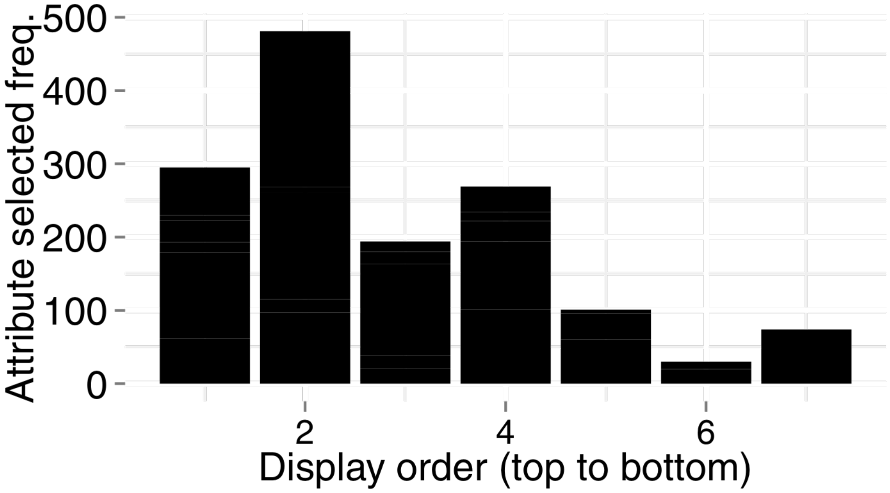


Figure 34: Raters were more likely to choose attributes displayed earlier on the page.

Opportunities for early feedback

To explore the possibility of automatic, early feedback, we trained a classifier using etcml.com to detect the most common errors for each question (Table 8). Because students unlikely to revise work without external feedback[158], even somewhat unreliable feedback (e.g., "Check to see that…") may have benefits.

Identify-verify uses its auto-graders confidence to indicate ambiguity. Might students benefit from knowing that peers may have trouble understanding them? Evidence from automated essay scoring suggests that well-designed early feedback may help students write clearer answers [159], [160].

Coping with fewer graders than submitters

Table 8: Algorithmically predicting errors could automate early feedback.

Attribute	Accuracy	Precision	Recall
Incorrect attribute: "The question assumes that the user has feelings about the feature" (Q1)	0.79	0.58	0.41
Missed attribute: "More individual exploration in the space of designs" (Q2)	0.59	0.64	0.79
Incorrect attribute: "Other incorrect/irrelevant answer" (Q2)	0.9	0.27	0.73

In Experiment 1, almost twice as many students submitted work as performed assessment; the rest dropped the class in the meanwhile. Experiment 2 was conducted later in the course, and a much larger fraction of the 850 students who submitted answers also assessed. Intelligently rationing raters is important in large online systems with voluntary participation. Identify-verify system handles this problem by rationing fewer graders for unambiguous answers. Because of the smaller number of raters, the system asked a median of only one identification per question, saving more identifications for the most ambiguous answers. For this experimental system, students were not penalized for not participating in assessment. Future work could explore penalties for non-participation, or incent assessment in other ways.

When should instructors use hybrid grading?

Peer assessment works best when staff spot-grade some student submissions because it helps staff refine assessment materials and baseline peer grades [28], [94]. However, courses may not have the resources for staff to grade several hundred examples that can train a machine-learning algorithm. (Even if it enables richer questions.)

Furthermore, requiring large amounts of training data may dissuade instructors from revising questions. We see two opportunities exist for future work. First, an online-learning algorithm may improve prediction accuracy as students assess each other. However, because the system would demand fewer assessments as its prediction accuracy increases, this may encourage free-riding. Future work could leverage such algorithms, while balancing for fairness. More immediately, assessment data from peers may be used to train algorithms. For example, an advanced cohort takes the class a week ahead of the general class. There are many exciting opportunities for integrating peer and algorithmic assessment to increase student learning and leverage the rater's time better.

Future work and Conclusion

This chapter demonstrated the feasibility of combining machine and peer grading through the identify-verify workflow. It showed how this workflow results in more detailed student feedback, and can be leveraged to provide early feedback. To further instructor experimentation and research, our open-source code is available at https://github.com/StanfordHCI/peerstudio. In addition, a hosted version of the platform is available at http://www.peerstudio.org.

Future work falls in three categories: First, this chapter assumes the final grade for a short-answer response can be expressed as a summed combination. Deploying this workflow in other classes may suggest other ways to structure assessment and verification, for e.g., as a decision tree. Second, many techniques in this chapter may be extended with algorithmic improvements. For instance, our system currently implements a fixed-control method for dynamically controlling the number of peer raters for a submission. A decision-theoretic model may result in even lower grading burden [150]. Similarly, an online learning algorithm could dynamically update estimates of the predicted grade to guide which ratings are collected [161]. Third, in this chapter, the system decided which answers a rater should assess and which assessments to verify based on what information was most valuable to determine the final grade. Because performing peer assessment is a valuable learning activity [94], future work may select submissions for raters that optimize both score/feedback quality and student learning (e.g. by choosing submissions for peer raters that they can learn most from).

We propose that the combination of machine and human grading can offer strengths that neither has in isolation. The large scale of online classes enables machines to effectively improve the educational experience [162]. By lessening grading burden, machines can focus peers on providing more detailed feedback. Automatic feedback may also focus students on topics they have not fully mastered. Likewise, peers can help machines identify “unknown unknowns” that are blind spots in their models, and help bootstrap that model quickly. Hybrid peer-machine approaches may also help in-person classes and many social computing areas, including crowdsourcing.

Acknowledgments

We thank Zhenghao Chen and Brennan Saeta at Coursera for building systems that enabled our experimental grading; Kathryn Papadopoulos, Lalida Sritanyaratana and community TAs for grading student submissions; Kanit (Ham) Wongsuphasawat for helping run the pilot experiment, and students that enrolled in our class and participated in this experiment.

Chapter 5
Rapid feedback for revision[7]

The power of rapid feedback

Online learning need not be a loop of watching video lectures and then submitting assignments. To most effectively develop mastery, students must repeatedly revise based on *immediate, focused feedback* [18]. Revision is central to the method of deliberate practice as well as to mastery learning, and depends crucially on rapid formative assessment and applying corrective feedback [163]. In domains as diverse as writing, programming, and art, immediate feedback reliably improves learning; delaying feedback reduces its benefit [164].

Unfortunately, many learning experiences cannot offer tight feedback-revision loops. When courses assign open-ended work such as essays or projects, it can easily take a week after submission to receive feedback from peers or overworked instructors. Feedback is also often coupled with an unchangeable grade, and classes move to new topics faster than feedback arrives. The result is that many opportunities to develop mastery and expertise are lost, as students have few opportunities to revise work and no incentive to do so.

Could software systems enable peers in massive classes to provide rapid feedback on in-progress work? In massive classes, peer assessment already provides summative grades and critiques on final work [37], but this process takes days, and is often as slow as in-person classes. This chapter instead introduces a peer learning design tailored for near-immediate peer feedback. It capitalizes on the scale of massive classes to connect students to trade structured feedback on drafts. This process can provide

[7] A version of this chapter was originally published as an article in the proceeding of the ACM Conference on Learning at Scale, 2015 as [291].

feedback to students within minutes of submission, and can be repeated as often as desired.

We present the *PeerStudio* system for fast feedback on in-progress open-ended work. Students submit an assignment draft whenever they want feedback and then provide rubric-based feedback on two others' drafts in order to unlock their own results. PeerStudio explicitly encourages mastery by allowing students to revise their work multiple times.

Even with the scale of massive classes, there are not always enough students online to guarantee fast feedback. Therefore, PeerStudio recruits students who are online already, and also those who have recently submitted drafts for review but are no longer online. PeerStudio uses a progressive recruitment algorithm to minimize the number of students emailed. It reaches out to more and more students, emailing a small fraction of those who recently submitted drafts each time, and stops recruiting immediately when enough (e.g., two) reviewers have been recruited.

This chapter reports on PeerStudio's use in two massive online classes and two in-person classes. In a MOOC where 472 students used PeerStudio, reviewers were recruited within minutes (median wait time: seven minutes), and the first feedback was completed soon after (median wait time: 20 minutes). Students in the two, smaller, in-person classes received feedback in about an hour on average. Students took advantage of PeerStudio to submit full drafts ahead of the deadline, and paid particular attention to free-text feedback beyond the explicit rubric.

A controlled experiment measured the benefits of rapid feedback. This between-subjects experiment assigned participants in a MOOC to one of three groups. One control group saw no feedback on in-progress work. A second group received feedback on in-progress work 24 hours after submission. A final group received feedback as soon as it was available. Students who received fast in-progress feedback had higher final grades than the control group ($t(98)=2.1$, $p<0.05$). The speed of the feedback was critical: receiving slow feedback was statistically indistinguishable from receiving no feedback at all ($t(98)=1.07$, $p=0.28$).

PeerStudio demonstrates how massive online classes can be designed to provide feedback an order of magnitude faster than many in-person classes. It also shows how MOOC-inspired learning techniques can *scale down* to in-person classes. In this case, designing and testing systems iteratively in massive online classes led to techniques that worked well in offline classrooms as well; Wizard of Oz prototyping and experiments in small classes led to designs that work well at scale. Finally, parallel deployments at different scales help us refocus our efforts on creating systems that produce pedagogical benefits at any scale.

Related work

PeerStudio relies on peers to provide feedback. Prior work shows peer-based critique is effective both for in-person [25], [94] and online classes [37], and can provide students accurate numeric grades and comments [37], [93].

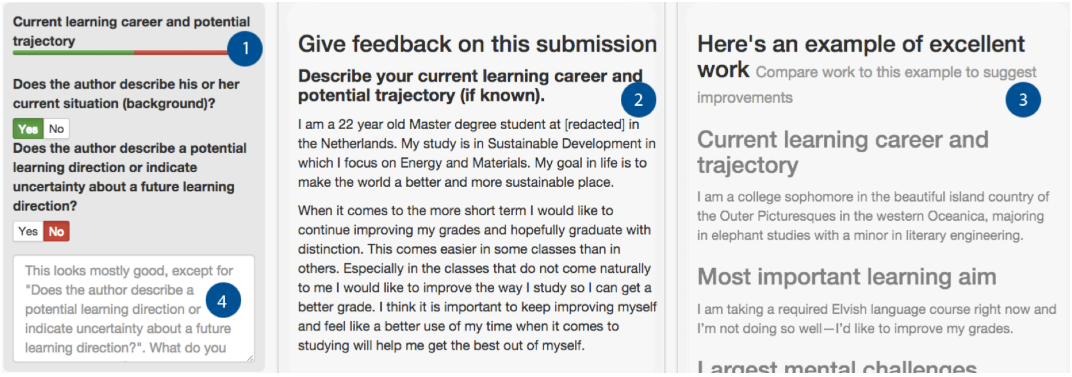
PeerStudio bases its design of peer feedback on prior work about how feedback affects learning. By *feedback*, we mean task-related information that helps students improve their performance. Feedback improves performance by changing students' locus of attention, focusing them on productive aspects of their work [165]. It can do so by making the difference between current and desired performance more salient [166], by explaining the cause of poor performance [167], or by encouraging students to use a different or higher standard to compare their work against [168].

Fast feedback improves performance by making the difference between the desired and current performance more salient [164]. When students receive feedback quickly (e.g., in an hour), they apply the concepts they learn more successfully [164]. In domains like mathematics, computers can generate feedback instantly, and combining such formative feedback with revision improves grades [169]. PeerStudio extends fast feedback to domains such as design and writing where automated feedback is limited and human judgment is necessary.

Feedback merely changes what students attend to, so not all feedback is useful, and some feedback degrades performance [165]. For instance, praise is frequently ineffective because it shifts attention *away* from the task and onto the self [170].

Therefore, feedback systems and curricular designers must match feedback to instructional goals. Large-scale meta-analyses suggest that the most effective feedback helps students set goals for future attempts, provides information about the quality of their current work, and helps them gauge whether they are moving towards a good answer [165]. Therefore, PeerStudio provides a low-cost way of specifying goals when students revise, uses a standardized rubric and free-form comments for correctness feedback, and a way to browse feedback on previous revisions for velocity.

How can peers provide the most accurate feedback? Disaggregation can be an important tool: summing individual scores for components of good writing (e.g. grammar and argumentation) can capture the overall quality of an essay more accurately than asking for a single writing score [171], [172]. Therefore, PeerStudio asks for individual judgments with yes/no or scale questions, and not aggregate scores.



Current learning career and potential trajectory

Does the author describe his or her current situation (background)?

[✓ Yes] No

Does the author describe a potential learning direction or indicate uncertainty about a future learning direction?

Yes [✓ No]

This looks mostly good, except for "Does the author describe a potential learning direction or indicate uncertainty about a future learning direction?". What do you

Give feedback on this submission

Describe your current learning career and potential trajectory (if known).

I am a 22 year old Master degree student at [redacted] in the Netherlands. My study is in Sustainable Development in which I focus on Energy and Materials. My goal in life is to make the world a better and more sustainable place.

When it comes to the more short term I would like to continue improving my grades and hopefully graduate with distinction. This comes easier in some classes than in others. Especially in the classes that do not come naturally to me I would like to improve the way I study so I can get a better grade. I think it is important to keep improving myself and feel like a better use of my time when it comes to studying will help me get the best out of myself.

Here's an example of excellent work Compare work to this example to suggest improvements

Current learning career and trajectory

I am a college sophomore in the beautiful island country of the Outer Picturesques in the western Oceanica, majoring in elephant studies with a minor in literary engineering.

Most important learning aim

I am taking a required Elvish language course right now and I'm not doing so well—I'd like to improve my grades.

Largest mental challenges

Figure 35: PeerStudio is a peer learning platform for rapid, rubric-based feedback on drafts. The reviewing interface above shows (1) the rubric, (2) the student draft, (3) an example of excellent work to compare student work against. PeerStudio scaffolds reviewers with automatically generated commenting tips (4).

PeerStudio uses the large scale of the online classroom in order to quickly recruit reviewers after students submit in-progress work. In contrast, most prior work has capitalized on scale only after all assignments are submitted. For instance, DeduceIt uses the semantic similarity between student solutions to provide automatic hinting and to check solution correctness [162], while other systems cluster solutions to help teachers provide feedback quickly [63].

Fast peer feedback with PeerStudio

Students can use PeerStudio to create and receive feedback on any number of drafts for every open-ended assignment. Because grades shift students' attention away from the task to the self [165], grades are withheld until the final version.

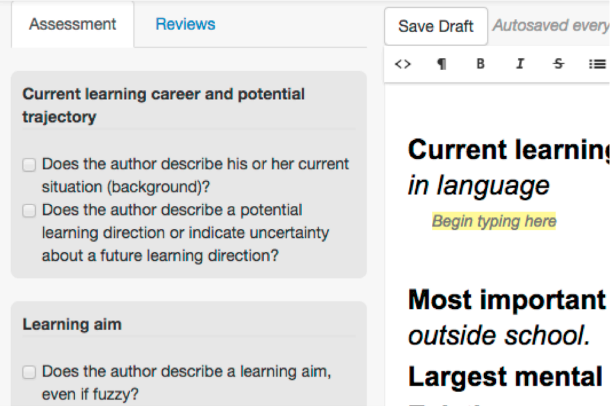
Creating a draft, and seeking feedback

PeerStudio encourages students to seek feedback on an initial draft as early as possible. When students create their first draft for an assignment, PeerStudio shows them a minimal, instructor-provided starter template that students can modify or overwrite (Figure 36). Using a template provides a natural hint for when to seek feedback—when the template is filled out. It also provides structure to students that need it, without constraining those who don't. To keep students focused on the most important aspects of their work, students always see the instructor-provided assignment rubric in the drafting interface (Figure 36, left). Rubrics in PeerStudio comprise a number of criteria for quality along multiple dimensions.

Students can seek feedback on their current draft at any time. They can focus their reviewers' attention by leaving a note about the kind of feedback they want. When students submit their draft, PeerStudio starts finding peer reviewers. Simultaneously, it invites the student to review others' work.

Reviewing peer work

PeerStudio uses the temporal overlap between students to provide fast feedback. When a student submits their draft, PeerStudio asks them to review their peers' submissions in order to unlock their own feedback [173]. Since their own work remains strongly activated, reviewing peer work immediately encourages students to reflect [174].



Current learning career and potential trajectory

□ Does the author describe his or her current situation (background)?
□ Does the author describe a potential learning direction or indicate uncertainty about a future learning direction?

Learning aim

□ Does the author describe a learning aim, even if fuzzy?

Current learnin
in language
Begin typing here

Most important
outside school.
Largest mental

Figure 36: The drafting interface shows the assignment rubric, and a starter template. Reviews on previous versions are also available (tab, top-left).

Students need to review two drafts before they see feedback on their work. Reviewing is double blind. Reviewers see their peer's work, student's review request notes, the instructor-created feedback rubric, and an example of excellent work to compare against. Reviewers' primary task is to work their way down the feedback rubric, answering each question. Rubric items are all yes/no or scale responses. Each group of rubric items also contains a free-text comment box, and reviewers are encouraged to write textual comments. To help reviewers write useful comments, PeerStudio prompts them with dynamically generated suggestions.

Reading reviews and revising

PeerStudio encourages rapid revision by notifying students via email immediately after a classmate reviews their work. To enable feedback comparison, PeerStudio displays the number of reviewers agreeing on each rubric question, as well as reviewers' comments. Recall that to emphasize iterative improvement, PeerStudio does not display grades, except for final work.

After students read reviews, PeerStudio invites them to revise their draft. Since reflection and goal setting are an important part of deliberate practice, PeerStudio asks students to first explicitly write down what they learned from their reviews and what they plan to do next.

PeerStudio also uses peer assessment for final grading. Students can revise their draft any number of times before they submit a final version to be graded. The final reviewing process for graded submissions is identical to early drafts, and reviewers see the same rubric items. For the final draft, PeerStudio calculates a grade as a weighted sum of rubric items from reviews for that draft.

PeerStudio integrates with MOOC platforms through LTI, which allows students to login using MOOC credentials, and automatically returns grades to class management software. It can be also used as a stand-alone tool.

PeerStudio design

PeerStudio's feedback design relies on rubrics, textual comments, and the ability to recruit reviewers quickly. We outline the design of each.

Rubrics

Rubrics effectively provide students feedback on the current state of their work for many open-ended assignments, such as writing [8], [9], design [37], and art [25]. Rubrics comprise multiple dimensions, with cells describing increasing quality along each. For each dimension, reviewers select the cell that most closely describes the submission; in between values and gradations within cells are often possible. Comparing and matching descriptions encourages raters to build a mental model of each dimension that makes rating faster and cognitively more efficient [175].

When rubric cell descriptions are complex, novice raters can develop mental models that stray significantly from the rubric standard, even if it is shown prominently [172]. To mitigate the challenges of multi-attribute matching, PeerStudio asks instructors to list multiple distinct criteria of quality along each dimension (Figure 37). Raters then explicitly choose which criteria are present. Criteria can be binary *e.g.*, "did the student choose a relevant quote that logically supports their opinion?” or scales, *e.g.*, “How many people did the student interview?”

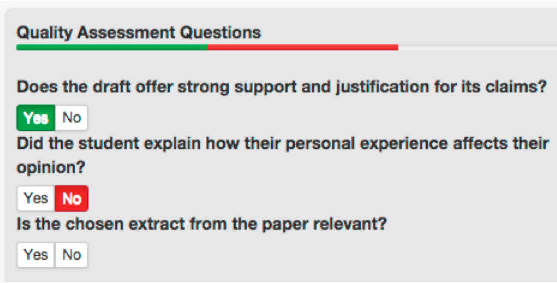
Our initial experiments and prior work suggest that given a set of criteria, raters satisfice by marking some but not all matching criteria [176]. To address this, PeerStudio displays binary questions as dichotomous choices, so students must choose either yes/no (Figure 37); and ensures that students answer scale questions by explicitly setting a value.

To calculate final grades, PeerStudio awards credit to yes/no criteria if a majority of reviewers marked it as present. To reduce the effect of outlying ratings, scale questions are given the median score of reviewers. The total assignment grade is the sum of grades across all rubric questions.

Scaffolding comments

Rubrics help students understand the current quality of their work; free-text comments from peers help them improve it. Reviews with accurate rubric scores, but without comments may provide students too little information.

To scaffold reviewers, PeerStudio shows short tips for writing comments just below the comment box. For instance, if the comment merely praises the submission and has no constructive feedback, it may remind students “Quick check: Is your feedback actionable? Are you expressing yourself succinctly?” Or it may ask reviewers to “Say more…” when they write “great job!”



Quality Assessment Questions

Does the draft offer strong support and justification for its claims?

[✓ Yes] No

Did the student explain how their personal experience affects their opinion?

Yes [✓ No]

Is the chosen extract from the paper relevant?

Yes No

Figure 37: Example dichotomous questions in PeerStudio. The last question is not yet answered. Students must choose yes/no before they can submit the review.

To generate such feedback, PeerStudio compiles a list of relevant words from the student draft and the assignment description. For example, for a critique on a research paper, words like “contribution”, “argument”, “author” are relevant. PeerStudio then counts the number of relevant words a comment contains. Using this count, and the comment’s total length, it suggests improvements. This simple heuristic catches a large number of low-quality comments. Similar systems have been used to judge the quality of product reviews online [177].

PeerStudio also helps students provide feedback that’s most relevant to the current state of the draft, by internally calculating the reviewer’s score for the submission. For a low-quality draft, it asks the reviewer, “What’s the first thing you’d suggest to get started?” For middling drafts, reviewers are asked, “This looks mostly good, except for [question with a low score]. What do you suggest they try?” Together, these commenting guides result in reviewers leaving substantive comments.

Recruiting reviewers

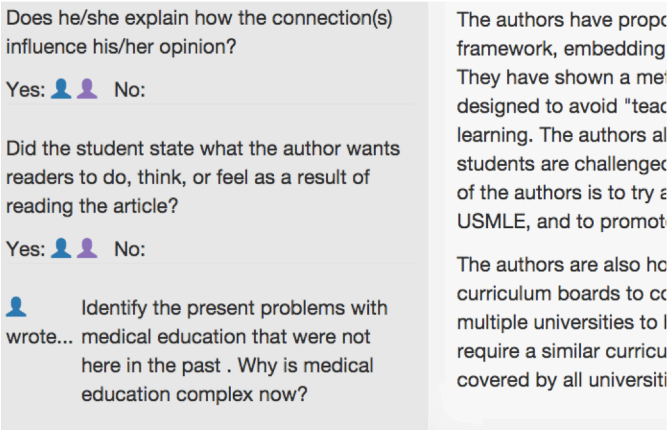
Because students review immediately after submitting, reviewers are found quickly when there are many students submitting one after another, *e.g.*, in a popular time zone. However, students who submit at an unpopular time still need feedback quickly. When enough reviewers are not online, PeerStudio progressively emails and enlists help from more and more students who have yet to complete their required two reviews, and enthusiastic students who have reviewed even before submitting a draft. PeerStudio emails a random selection of five such students every half hour, making sure the same student is not picked twice in a 24-hour period. PeerStudio stops emailing students when all submissions have at least one review. This enables students to quickly receive feedback from one reviewer and begin revising.

To decide which submissions to show reviewers, PeerStudio uses a priority queue. This queue prioritizes student submissions by the number of reviews (submissions with the fewest, or no, reviews have highest priority), and by the time the submission has been in the review queue. The latest submissions have the highest priority. PeerStudio seeks two reviewers per draft.

Field Deployment: in-person and at scale

This chapter describes PeerStudio deployments in two open online classes: *Learning How to Learn* (603 students submitting assignments), *Medical Education in the New Millennium* (103 students) on the Coursera and OpenEdX platforms respectively. We also describe deployments in two in-person classes: a senior-level class at the University of Illinois at Urbana-Champaign on *Social Visualization* (125 students), and a graduate-level class in education at Stanford University, on *Technology for Learners* (51 students).

All four classes used PeerStudio for open-ended writing assignments. In *Learning how to Learn*, for their first assignment students wrote an essay about a learning goal and how they planned to accomplish it using what they learned in class (*e.g.*, one student wrote about being "an older student in Northern Virginia retooling for a career in GIS after being laid off"). In the second assignment, they created a portfolio, blog or website to explain what they learned to others (*e.g.*, one wrote: "I am a professor of English as a Second Language at a community college. I have created a PowerPoint presentation for my colleagues [about spaced repetition and frequent testing]").



Does he/she explain how the connection(s) influence his/her opinion?

Yes: 👤👤 No:

Did the student state what the author wants readers to do, think, or feel as a result of reading the article?

Yes: 👤👤 No:

👤 wrote... Identify the present problems with medical education that were not here in the past . Why is medical education complex now?

The authors have propc
framework, embedding
They have shown a me
designed to avoid "teac
learning. The authors al
students are challengec
of the authors is to try a
USMLE, and to promot

The authors are also ho
curriculum boards to cc
multiple universities to l
require a similar curricu
covered by all universiti

Figure 38: Students see reviews in the context of their draft (right, clipped). PeerStudio displays the number of reviewers (two here) agreeing on each rubric question and comments from each.

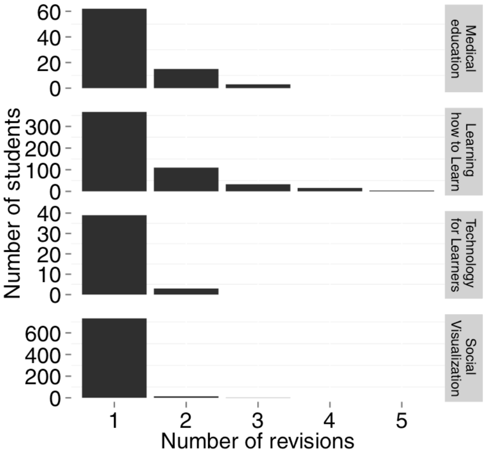


Figure 39: Most students created a single revision. Students in MOOCs revised more than students in in-person classes.

The Social Visualization and Medical Education classes asked students to critique research papers in the area. In Social Visualization, students also used PeerStudio for an open-ended design project on data visualization (*e.g.*, one student team designed a visualization system that used data from Twitter to show crisis needs around the US). Finally, the Technology for Learners class used PeerStudio as a way to critique a learning tool (*e.g.*, ClassDojo, a classroom discipline tool). This class requested its reviewers to sign reviews, so students could follow-up with each other for lingering questions.

Deployment observations

Throughout these deployments, we read students' drafts, feedback, and revisions. We regularly surveyed students about their experiences, and spoke to instructors about their perspectives. Several themes emerged.

Students requested feedback on full rough drafts

Rather than submit sections of drafts, students submitted full rough drafts. Drafts were often missing details (*e.g.,* lacking examples). In the Medical Education critique, one question was "did you find yourself mostly agreeing or mostly disagreeing with the content of the research paper? Why?" In initial drafts, students often pointed out only

one area of disagreement, later drafts added the rest. Other drafts were poorly explained (*e.g.,* lacking justification for claims) or too rambling.
Students typically asked for four kinds of feedback: 1) On a specific aspect of their work, *e.g.*, "I guess I need help with my writing, vocabulary and grammar, since I'm not an English native-speaker"; 2) On a specific component of the assignment: *e.g.*, "Can you let me know if part 4 and 5 make sense—I feel like I am trying to say too much all in one go!" 3) As a final check before they turned in their work: *e.g.*, "This draft is actually a 'release candidate'. I would like to know if I addressed the points or if I missed something." 4) As a way to connect with classmates: *e.g.*, "I just want to hear about your opinions :)".

When students revised their draft, we asked, "Overall, did you get useful feedback on your draft?" as a binary question—80% answered 'yes'.

Students revise rarely, especially in in-person classes

Most students did not create multiple drafts (Figure 39). Students in the two MOOCs were more likely to revise than students in in-person classes (t(1404)=12.84, p< 0.001). Overall, 30.1% of online students created multiple revisions, but only 7% of those in in-person classes did.

When we asked TAs in the in-person classes why so few students revised, they told us they did not emphasize this feature of PeerStudio in class. Furthermore, student responses in surveys indicated that many felt their schedule was too busy to revise. One wrote it was unfair to "expect us to read some forty page essays, then write the critiques and then review two other people, and then make changes on our work... twice a week." These comments underscore that merely creating software systems for iterative feedback is not enough—an iterative approach must be reflected in the pedagogy as well.

Students see comments as more useful than rubric feedback

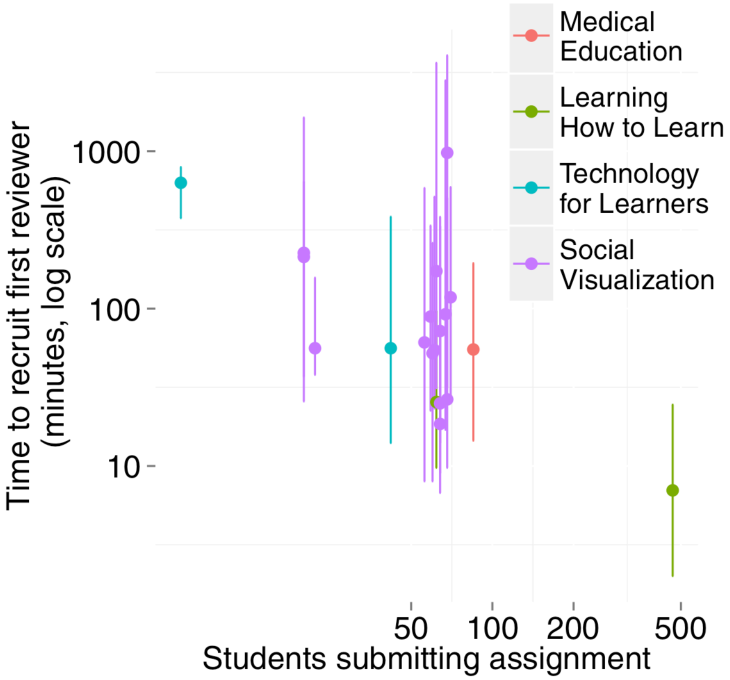


Figure 40: Reviewers are recruited faster in larger classes.

Students could optionally rate reviews after reading them and leave comments to staff. Students rated 758 of 3,963 reviews. We looked at a random subset of 50 such comments. In their responses, students wrote that freeform comments were useful (21 responses) more often than rubric-based feedback (5 responses). Students also disagreed more with reviewers' comments (7 responses) than with their reviewers' marked rubric (3 responses). This is possibly because comments can capture useful interpretive feedback, but differences in interpretation lead to disagreement.

An undergraduate TA looked at a random subset of 150 student submissions, and rated reviewer comments on a 7-point Likert scale on how concretely they helped students revise. For example, here is a comment that was rated "very concrete (7)" on an essay about planning for learning goals:

"What do you mean by 'good schedule'? There's obviously more than one answer to that question, but the goal should be to really focus and narrow it down. Break a larger goal like "getting a good schedule" into concrete steps such as: 1) get eight hours of sleep, 2)…

We found 45% of comments were “somewhat concrete” (a rating of 5 on the scale) or better, and contained pointers to resources or specific suggestions on how to improve; the rest of the comments were praise or encouragement. Interestingly, using the same 7-point Likert scale, students rated reviews as concrete more often than the TA (55% of the time).

Students reported relying on comments for revising. For instance, the student who received the above comment wrote, “I somehow knew I wasn't being specific… The reviewer's ideas really helped there!” The lack of comments was lamented upon, “The reviewer did not comment any feedback, so I don’t know what to do.”

One exception to the general trend of comments being more important was students who submitted ‘release candidate’ drafts for a final check. Such students relied heavily on rubric feedback: “I have corrected every item that needed attention to. I now have received all yes to each question. Thanks guys. :-)”

Comments encourage students to revise

The odds of students revising their drafts increase by 1.10 if they receive any reviews with free-form comments ($z=4.6$, $p<0.001$). Since fewer than half the comments contained specific improvement suggestions, this suggests that, in addition to being informational, reviewer comments also play an important motivational role.

Revisions locally add information, improve understandability

We looked at the 100 reflections that students wrote while starting the revision to understand what changes they wanted to make. A majority of students (51%) intended to add information based on their comments, *e.g.*, “The math teacher [one of the reviewers] helped me look for other sources relating to how math can be fun and creative instead of it being dull!” A smaller number (16%) wanted to change how they had expressed ideas to make them easier to understand, *e.g.*, “I did not explain clearly the three first parts… I shall be clearer in my re-submission” and, “I do need to avoid repetition. Bullets are always good.” Other changes included formatting, grammar, and occasionally wanting a fresh start. The large fraction of students who wanted to add information to drafts they previously thought were complete suggests that peer feedback helps students see flaws in their work, and provides new perspectives.

Most students reworked their drafts as planned: 44% of students made substantive changes based on feedback, 10% made substantive changes not based on the comments received, and the rest only changed spelling and formatting. Most students added information to or otherwise revised one section, while leaving the rest unchanged.

PeerStudio recruits reviewers rapidly

We looked at the PeerStudio logs to understand the platform's feedback latency. Reviewers were recruited rapidly for both in in-person and online classes (see Figure 40), but the scale of online classes has a dramatic effect. With just 472 students using the system for the first assignment in Learning How to Learn, the median recruitment-time was 7 minutes and the 75th quartile was 24 minutes.

Few students have long wait times

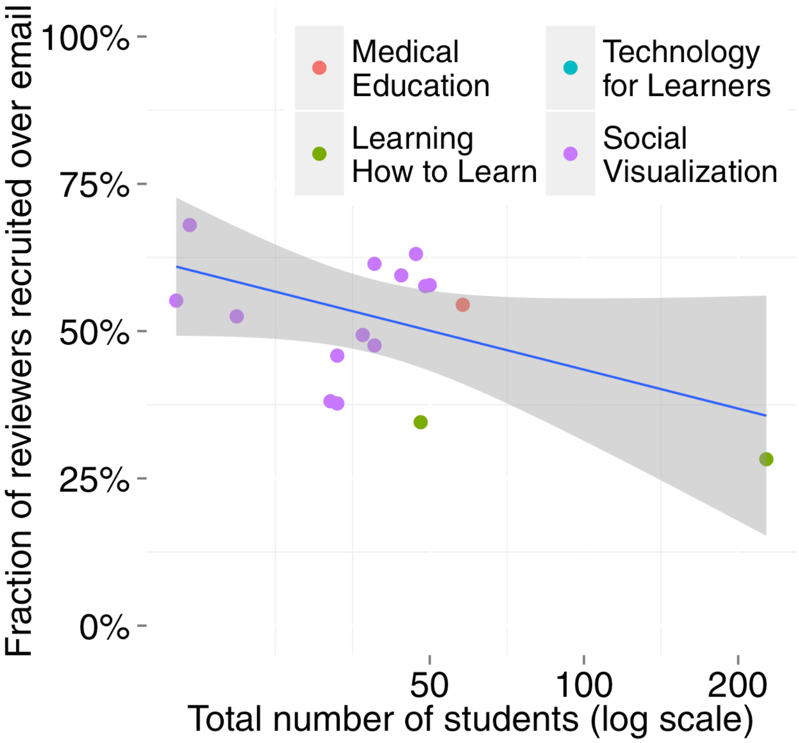


Figure 41: More students in large classes are likely to be online at the same time, so fewer reviewers were recruited by email.

PeerStudio uses a priority queue to seek reviews; it prioritizes newer submissions given two submissions with the same number of reviews. This reduces the wait time for the *average* student, but unlucky students have to wait longer (e.g. when they submit just before a popular time, and others keep submitting newer drafts). Still, significant delays are rare: 4.4% had no reviews in the first 8 hours; 1.8% had no reviews in 24 hours. To help students revise, staff reviewed submissions with no reviews after 24 hours.

Feedback latency is consistent even early in the assignment

Even though fewer students use the website farther from the deadline, peer review means that the workload and review labor automatically scale together. We found no statistical difference in recruitment time (t(1191)= 0.52, p=0.6) between the first two and last two days of the assignment, perhaps because PeerStudio uses email to recruit reviewers.

Fewer reviewers recruited over email with larger class size.

PeerStudio emails students to recruit reviewers only when enough students aren't already on the website. In the smallest class with 46 students submitting, 21% of reviews came from Web solicitation and 79% of reviews were written in response to an emailed request. In the largest, with 472 students submitting, 72% of reviews came from Web solicitation and only 28% from email (Figure 41). Overall, students responded to email requests approximately 17% of the time, independent of class size. These results suggest that PeerStudio achieves quick reviewing in small, in-person classes by actively bringing students online via email, and that this becomes less important with increasing class size, as students have a naturally overlapping presence on site.

Reviewers spend about ten minutes per draft

PeerStudio records the time between when reviewers start a review and when they submit it. In all classes except the graduate level Technology for Learners, students spent around 10 minutes reviewing each draft (Figure 43). The median reviewer in the graduate Technology for Learners class spent 22 minutes per draft. Because all students in that class started reviewing in-class but finished later, its variance in reviewing times is also much larger.

Are reviewers accurate?

There is very strong agreement between individual raters while using the rubric. In online classes, the median pair-wise agreement between reviewers on a rubric question is 74%, while for in-person classes it is 93%. However, because most drafts completed a majority of the rubric items successfully, baseline agreement is high, so Fleiss' κ is low. The median κ=0.19 for in-person classes, and 0.33 for online classes, conventionally considered "Fair agreement". In in-person classes, on average staff and students agreed on rubric questions 96% of the time.

Staff and peers write comments of similar length

Both in-person and online, the median comment was 30 words long (Figure 42). This

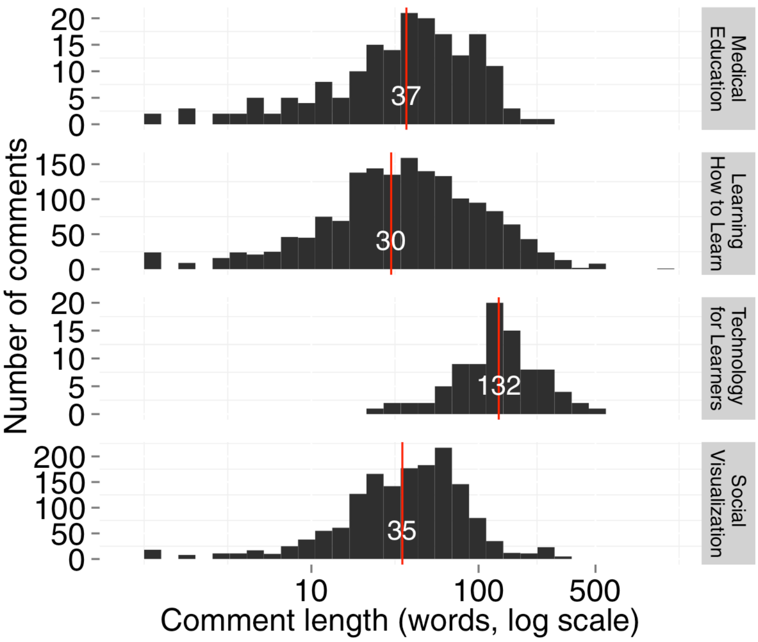


Figure 42: Students write substantive comments, both in-person and online. The graduate level Technology for Learners has longer comments, possibly because reviews were signed.

length compares well with staff comments in the *Social Visualization* class, which had a median of 35 words. Most reviews (88%) had at least some textual comments, in addition to rubric-based feedback.

Students trade-off reviewing and revising

23% of students reviewed more than the required two drafts. Survey results indicated that many such students used reviewing as an inexpensive way to make progress on their own draft. One student wrote that in comparison to revising their own work, "being able to see what others have written by reviewing their work is a better way to get feedback." Other students reviewed peers simply because they found their work interesting. When told she had reviewed 29 more drafts than required, one student wrote, "I wouldn't have suspected that. I kept reading and reviewing because people's stories are so interesting."

Students appreciate reading others' work more than early feedback and revision

A post-class survey in *Technology For Learners* asked students what they liked most about PeerStudio (30 responses). Students most commonly mentioned (in 13 responses) interface elements such as being able to see examples and rubrics. Reading each other's work was also popular (8 responses), but the ability to revise was rarely mentioned (3 responses). This is not surprising, since few students revised work in in-person classes.

Apart from specific usability concerns, students' most frequent complaint was that PeerStudio sent them too much email. One wrote, "My understanding was that students would receive about three, but over the last few days, I've gotten more." Currently, PeerStudio limits how frequently it emails students; future work could also limit the total number of emails a student receives.

Field Experiment: Does fast feedback on in-progress work improve final work?

The prior study demonstrated how students solicited feedback and revised work, and how quickly they can obtain feedback. Next, we describe a field experiment that asks two research questions: First, does feedback on in-progress work improve student performance? Second, does the speed of feedback matter? Do students perform better if they receive rapid feedback? We conducted this controlled experiment in *ANES 204: Medical Education in the New Millennium*, a MOOC on the OpenEdX platform. Students in this class had working experience in healthcare professions, such as medical residents, nurses and doctors. In the open-ended assignment, students read and critiqued a recent research paper based on their experience in the healthcare field. For example, one critique prompt was "As you read, did you find yourself mostly agreeing or mostly disagreeing with the content? Write about three points from the article that justify your support or dissent." The class used PeerStudio to provide students both in-progress feedback and final grades.

Method

A between-subjects manipulation randomly assigned students to one of three conditions. In the *No Early Feedback* condition, students could only submit one final draft of their critique. This condition generally mimics the status quo in many classes, where students have no opportunities to revise drafts with feedback. In the *Slow Feedback* condition, students could submit any number of in-progress drafts, in addition to their final draft. Students received peer feedback on all drafts, but this feedback wasn't available until 24 hours after submission. Additionally, students were only emailed about their feedback at that time. This condition mimics a scenario where a class offers students the chance to revise, but is limited in its turnaround time due to limited staff time or office hours. Finally, in the *Fast Feedback* condition, students could submit drafts as in the slow feedback condition, but were shown reviews as soon as available, mirroring the standard PeerStudio setup.

Students in all conditions rated their peers' work anonymously; reviewers saw drafts from all conditions and rated them blind to condition. Our server introduced all delays for the Slow Feedback condition after submission. Rubrics and the interface students used for reviewing and editing were identical across conditions.

Measures

To measure performance, we used the grade on the final assignment submission as calculated by PeerStudio. Since rubrics only used dichotomous questions, each rubric question was given credit if a majority of raters marked "yes". The grade of each draft was the sum of credit across all rubric questions for that draft.

Participants

In all, 104 students participated. Of these, three students only submitted a blank essay; their results were discarded from analysis. To analyze results, we built an ordinary-least-squares regression model with the experimental condition as the predictor variable, using *No Early Feedback* as the baseline (R^2=0.02).

Manipulation check

While PeerStudio can provide students feedback quickly, this feedback is only useful if students actually read it. Therefore, we recorded the time students first read their feedback. The median participant in the *Fast Feedback* condition read their reviews 592 minutes (9.8 hours) after submission; the median for the Slow Feedback condition was 1528 minutes (26.6 hours). This suggests that the manipulation effectively delayed feedback, but the difference between conditions was more modest than planned.

Results: fast early feedback improves final grades

Students in the *Fast Feedback* condition did significantly better than those in *No Early Feedback* condition (t(98)=2.1, $p<0.05$). On average, students scored higher by 4.4% of the assignment's total grade: i.e., enough to boost a score from a B+ to an A-.

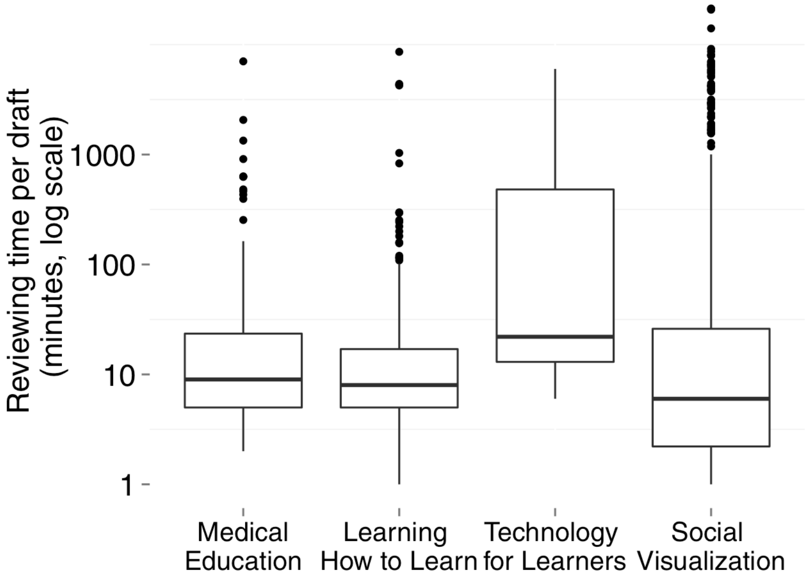


Figure 43: Reviewers spend roughly 10 minutes reviewing each draft. The graduate-level Technology for Learners class spends longer. (The larger variation is because students start reviewing in class, and finish later.)

Slow early feedback yields no significant improvement

Surprisingly, we found that students in the *Slow Feedback* condition did not do significantly better than those in the *No Early Feedback* condition (t(98)=1.07, *p=0.28*). These results suggest that for early feedback to improve student performance, it must be delivered quickly.

Because of the limited sample size, it is also possible this experiment was unable to detect the (smaller) benefits of delayed early feedback.

Students with fast feedback don't revise more often

There was no significant difference between the number of revisions students created in the *Fast* and *Slow feedback* conditions (t(77)=0.2, p=0.83): students created on average 1.33 drafts; only 22% of students created multiple revisions. On average, they added 83 words to their revision, and there was no significant difference in the quantity of words changed between conditions (t(23)=1.04, p=0.30).

However, students with *Fast feedback* referred to their reviews marginally more frequently when they entered reflections and planned changes in revision ($\chi^2(1)$=2.92, p=0.08). This is consistent with prior findings that speed improves performance by making feedback more salient.

Even with only a small number of students revising, the overall benefits of early feedback seem sizeable. Future work that better encourages students to revise may further increase these benefits.

Discussion

The field deployment and subsequent experiment demonstrate the value of helping students revise work with fast feedback. Even with a small fraction of students creating multiple revisions, the benefits of fast feedback are apparent. How could we design pedagogy to amplify these benefits?

Redesigning pedagogy to support revision and mastery

In-person classes are already using PeerStudio to change their pedagogy. These classes did not use PeerStudio as a way to reduce grading burden: both classes still had TAs grade every submission. Instead, they used PeerStudio to expose students to each other's work and to provide them feedback faster than staff could manage.

Fully exploiting this opportunity will require changes. Teachers will need to teach students about when and how to seek feedback. Currently, PeerStudio encourages students to fill out the starter template before they seek feedback. For some domains, it may be better to get feedback using an outline or sketch, so reviewers aren't distracted by superficial details[178]. In domains like design, it might be useful to get feedback on multiple alternative designs[179]. PeerStudio might explicitly allow these different kinds of submissions.

PeerStudio reduces the time to get feedback, but students still need time to work on revisions. Assignments must factor this revision time into their schedule. We find it heartening that 7% of in-person students actually revised their drafts, even when their assignment schedules were not designed to allow it. That 30% of online students revised assignments may partly be because schedules were designed around the assumption that learners with full-time jobs have limited time: consequently, online schedules often provide more time between assignment deadlines.

Finally, current practice rewards students for the final quality of their work. PeerStudio's revision process may allow other reward schemes. For instance, in domains like design where rapid iteration is prized [66], [180], classes may reward students for sustained improvement.

Plagiarism

Plagiarism is a potential risk of sharing in-progress work. While plagiarism is a concern with all peer assessment, it is especially important in PeerStudio because the system shares work before assignments are due. In classes that have used PeerStudio so far, we found one instance of plagiarism: a student reviewed another's essay and then submitted it as their own. While PeerStudio does not detect plagiarism currently, it does record what work a student reviewed, as well as every revision. This record can help instructors check that the work has a supporting paper trail. Future work could automate this.

Another risk is that student reviewers may attempt to fool PeerStudio by giving the same feedback to every assignment they review (to get past the reviewing hurdle quickly so they can see feedback on their work). We observed three such instances. However, 'shortcut reviewing' is often easy to catch with techniques such as inter-rater agreement scores [181].

Bridging the in-person and at-scale worlds

While it was designed for massive classes, PeerStudio "scales down" and brings affordances such as fast feedback to smaller in-person classes. PeerStudio primarily relies on the natural overlap between student schedules at larger scales, but this overlap still exists at smaller scale and can be augmented via email recruitment.

PeerStudio also demonstrates the benefits of experimenting in different settings in parallel. Large-scale between-subjects experiments often work better online than in-person because in-person, students are more likely to contaminate manipulations by communicating outside the system. In contrast, in-person experiments can often be run earlier in software development using lower-fidelity approaches and/or greater support. Also, it can be easier to gather rich qualitative and observational data in person, or modify pilot protocols on the fly. Finally, consonant results in in-person and online deployments lend more support for the fundamentals of the manipulation (as opposed to an accidental artifact of a deployment).

Future work

Some instructors we spoke to worried about the overhead that peer assessment entails (and chose not to use PeerStudio for this reason). If reviewers spend about 10 minutes reviewing work as in our deployment, peer assessment arguably incurs a 20-minute overhead per revision. On the other hand, student survey responses indicate that they found looking at other students' work to be the most valuable part of the assessment process. Future work could quantify the benefits of assessing peer work, including inspiration, and how it affects student revisions. Future work could also reduce the reviewing burden by using early reviewer agreement to hide some rubric items from later reviewers [172].

Matching reviewers and drafts

PeerStudio enables students to receive feedback from peers at any time, but their peers may be far earlier or more advanced in their completion of the assignment. Instead, it may be helpful to have drafts reviewed by students who are similarly advanced or just starting. Furthermore, students learn best from examples (peer work) if they are approachable in quality. In future work, the system could ask or learn the rough state of the assignment, and recruit reviewers who are similar.

Conclusion

This chapter suggests that the scale of massive online classes enables systems that drastically and reliably reduce the time to obtain feedback and creates a path to iteration, mastery and expertise. These advantages can also be scaled-down to in-person classrooms. In contrast to today's learn-and-submit model of online education, we believe that the continuous presence of peers holds the promise of a far more dynamic and iterative learning process.

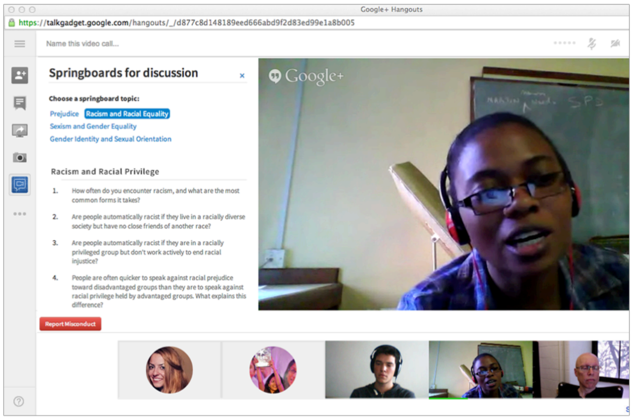
Chapter 6
Leveraging geographic diversity for classroom discussion

A version of this chapter was originally published as an article in the proceeding of the ACM Conference on Computer Supported Collaborative Work and Social Computing as [182].

Massive-scale diversity: an overlooked opportunity

At their best, culturally diverse classrooms leverage students' different backgrounds to improve learning and foster cultural understanding. When students engage with peers from different cultures, they become aware of their own assumptions and how others have different perspectives [183]. This shifts students from 'automatic' thinking to more 'active, effortful, conscious' thinking, which aids learning and growth [21]. But, while physical classrooms often strive to be diverse, they remain limited by physical geography [184].

Massive online courses recruit thousands of students from over 100 countries, bringing together peers with many nationalities and experiences [185]. Instructors often



Springboards for discussion

Choose a springboard topic:

Prejudice | Racism and Racial Equality | Sexism and Gender Equality | Gender Identity and Sexual Orientation

Racism and Racial Privilege

1. How often do you encounter racism, and what are the most common forms it takes?
2. Are people automatically racist if they live in a racially diverse society but have no close friends of another race?
3. Are people automatically racist if they are in a racially privileged group but don't work actively to end racial injustice?
4. People are often quicker to speak against racial prejudice toward disadvantaged groups than they are to speak against racial privilege held by advantaged groups. What explains this difference?

Google+ Hangouts
https://talkgadget.google.com/hangouts/_/d877c8d148189eed666abd9f2d83ed99e1a8b005
Name this video call...
Google+
Report Misconduct

Figure 44: Talkabout provides a structured discussion agenda and enables students from around the world to discuss with each other.

Course Title	Representative Discussion topics
Critical Perspectives on Management	How do you define innovation and invention? How do manage them? Are shipping containers and labor unions innovations or inventions?
Irrational Behavior	How do you treat money as a relative rather than absolute good? Do you think that it is more painful to pay with cash than credit? How might issues of fairness vary by culture?
Organizational Analysis	Describe your experience in organizations where decisions by organized anarchy occurred. Did they solve anything? How common were they?
Social Psychology	In your country, which forms of prejudice are the most socially acceptable, and which ones are the least acceptable? Why are some forms more acceptable than others?
Think Again	Since inductive arguments are defeasible, how can it ever be reasonable to trust them? Are arguments from analogy really different from inferences to the best explanation?

Table 9: Excerpts from discussion agendas from one week in different classes. Each question below included more detailed guidance in the actual discussion

advertise how many countries are represented in the class [36], [185], [186]. However, while student diversity has become a calling card of online education, this potential is currently untapped. Most online students currently see only a glimpse of their peers' global diversity, primarily in text discussion forums. This slow-motion communication is a poor fit for the open-ended dialogue characteristic of dorm hallway conversation [187], and can reinforce a one-size-fits-all, broadcast educational approach [188].

This chapter illustrates the potential of leveraging diversity in online classes, and introduces the Talkabout environment and curricula for small, geographically-diverse groups in massive classes. Talkabout connects students to their global peers via guided, synchronous video discussion. Talkabout focuses on harnessing *geographic diversity*, where students connect with peers from other parts of the world. Geographic diversity enables students to access peers with different cultures [189], levels of income [190], and beliefs about learning [191].

Geographically diverse classrooms can improve educational experiences, making them deeper and more realistic. Multinational discussions create the opportunity for what one student called a ‘mini United Nations’, where students experience first-hand the differing concerns and beliefs of people from different countries.

Talkabout forms groups of two to nine students from different parts of the world for a video discussion. Discussion prompts ask peers to relate course content to their local and personal experiences, encouraging students to reflect on previously unexamined assumptions about their own environments, and deepening their learning [192]. To date, more than 5,000 students from 134 countries have used Talk-about in fourteen online classes via Coursera and OpenEdX. This chapter reports results from the first seven courses and 3,200 students. These classes included Social Psychology, Organizational Analysis, Behavioral Economics, and Logic and Design. Table 9 shows a sampling of topics discussed. The median discussion had six students from five countries.

Talkabout’s discussion sessions improved student engagement: students randomly assigned to a Talkabout group were significantly more likely to participate in class quizzes than those placed on a waitlist for future participation (Wald z*=1.96, p=0.03).

Geographically diverse discussions yield higher grades and engagement. A controlled experiment in two massive online classes varied the number of countries present in Talkabout discussions. Students in more geographically diverse discussions performed significantly better on subsequent quizzes and exams (t(129)=1.78 and t(110)=2.03, *$p<0.05$*).

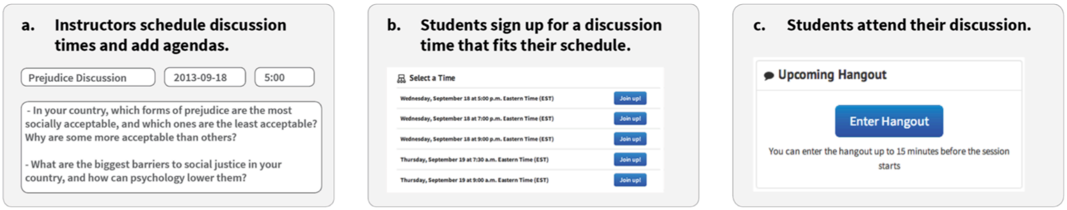
Some argue that online education is only desirable when face-to-face education is unavailable [193]. This chapter illustrates the benefits of inverting this proposition: global diversity enables online classrooms to create powerful, previously unavailable educational experiences and new forms of peer education at scale that go “beyond being there” [42].

Related Work

A tremendous benefit of diverse classrooms is that students of differing gender, ethnicity, and ability have opportunities to interact. When people interact with similar peers, their shared background leads to automatic thinking. In contrast, interacting with diverse peers often creates a discontinuity [21] that unearths hidden assumptions—yielding more active, effortful and conscious thought [194]. This active and effortful thinking improves academic performance and makes students more inclusive and democratic [21].

Travel, and interacting with geographically diverse people, similarly induces active thinking and reflection [192]. For instance, study-abroad programs result in deeper knowledge and understanding—especially about culture and international affairs—and greater self confidence [195].

The benefits of interacting with geographically diverse peers arise from differences in experiences and thinking. Examples of these differing experiences include stark differences in population density, income and educational systems [196]. People from different parts of the world have different cultural values, reasoning, and preferred learning methods. For instance, cultures differ in their emphasis of individuality versus interdependence [197], [198] and holistic versus analytical thinking [199]. These differences impact cognition. For example, when cultures encourage people to consider objects in relation with their context, they more often apply analogical thinking. By



a. Instructors schedule discussion times and add agendas.

Prejudice Discussion | 2013-09-18 | 5:00

- In your country, which forms of prejudice are the most socially acceptable, and which ones are the least acceptable? Why are some more acceptable than others?

- What are the biggest barriers to social justice in your country, and how can psychology lower them?

b. Students sign up for a discussion time that fits their schedule.

Select a Time

Wednesday, September 18 at 5:00 p.m. Eastern Time (EST)

Wednesday, September 18 at 7:00 p.m. Eastern Time (EST)

Wednesday, September 18 at 9:00 p.m. Eastern Time (EST)

Thursday, September 19 at 7:30 a.m. Eastern Time (EST)

Thursday, September 19 at 9:00 a.m. Eastern Time (EST)

c. Students attend their discussion.

Upcoming Hangout

You can enter the hangout up to 15 minutes before the session starts

Figure 45: Talkabout discussion timeline: (a) Instructors enter a discussion agenda, and times for the discussion. (b) Students pick their preferred time. (c) When they log on to Talkabout at their selected time, Talkabout assigns them to a group, and creates a private hangout. (c) Students show up at their selected time, and enter the discussion.

contrast, when people consider objects in isolation, they more often apply categorical rules [199].

To maximize the benefits of diversity, prior work emphasizes two factors: the numeric representation of diverse groups *(structural diversity);* and the number of settings that students interact in (*experiential diversity)* [49]. Ideally, students must meet frequently, and with equal status, in situations where collaboration is necessary and stereotypes are disconfirmed [50], and where differing views are welcomed [51].

Informed by this research, Talkabout forms geographically diverse discussion groups, and encourages fluid roles and consensus-based decisions that emphasize equality. Furthermore, Talkabout contributes a curriculum where students can question stereotypes and compare their views to their peers.

In most current online classes, students' opportunities for discussions with diverse peers are limited to text-based forums. Such asynchronous text channels inhibit trust-formation [200] and open-ended discussion [201]. Synchronous channels, such as video, improve participants' sense of belonging and willingness to collaborate [202]. Channels such as video which support multimodal communication and nonverbal cues are also better suited to ambiguous discussions [203] and complex sense-making [204]. For these reasons, Talkabout leverages synchronous, small-group video discussions to encourage meaningful, open-ended dialogue.

Massive scale presents both a formidable challenge and a powerful opportunity for online education. Prior work encouraging unstructured discussion failed to find an improvement in students' sense of community or academic achievement [205]. More systematically structured approaches have enjoyed greater success. One example is the use of rater redundancy and short exercises that create micro-expertise in peer review: with this structure, peers can provide expert-quality assessment and feedback [37], and act as mentors [206]. Talkabout introduces a structured interaction and curriculum that leverages diversity.

Coordinating global small-group discussion

The Talkabout interface guides instructors through setting up their course on Talkabout, and creating a structured discussion agenda for students (Figure 45a). This agenda is displayed throughout the discussion.

Students choose a discussion time from the published set (Figure 45b), up to a week in advance. As students log in at their selected time, Talkabout assigns them to groups (instructor can choose group size between 2 and 9). Talkabout has several policies for group assignment; by default it assigns arriving students to a group until it reaches its size limit; then it starts a new group. Other policies, discussed later, explicitly factor geographic location into group assignment. Discussions occur through the Google Hangouts platform for multi-person video and audio chat. For each group, Talkabout creates a discussion session exclusively for the assigned participants. Discussion groups exist only for the duration of the discussion session. If students participate in multiple discussion sessions—even in the same course and on the same topic—they are likely to have different partners, because grouping depends on students' arrival order. Consequently, students hear different ideas and experiences each time.

During discussions, the Talkabout Hangout application shows the instructor's discussion agenda on the left and the video chat on the right. An agenda typically includes suggested discussion topics or activities (Figure 44, Figure 47).

Assignment by arrival yields diverse groups

To quantify the geographic diversity in discussions, we aggregate countries into eight geographical regions, and count the number of regions in each discussion. Five regions are from the World Bank's classification [207]: *Eastern Europe and Central Asia* (primarily the former Soviet bloc), *East Asia and Pacific* (mainly China, Japan, Korea, and South-east Asia), *South Asia* (mainly the Indian subcontinent), *Latin America and the Caribbean* (Americas except the US and Canada), *Middle East and North Africa*, and *Sub-Saharan Africa*. The World Bank only classifies middle- and low-income countries, so we added three other regions: *North America* (US and Canada), *Western Europe*, and *South Pacific* (primarily Australia and Polynesia).

Across seven classes and the first 3,200 participants, allocating six-person groups by arrival order yielded discussions with a median of four global regions (Figure 46b), and a median of five countries (Figure 46a). The median pair-wise distance between discussants was approx. 6,600km (4,100 mi): more than the distance between New York and London.

Structuring Talkabout discussions

Our early experiences with Talkabout, as well as prior work, suggest that it is critical to co-design curricular strategies with educational interaction design. In particular,

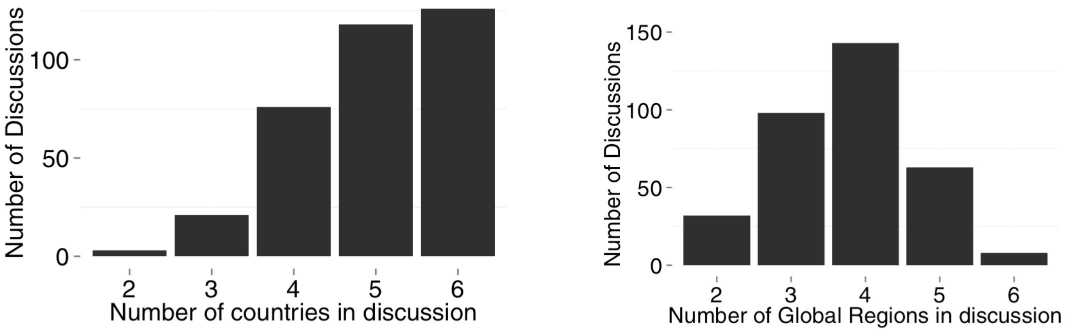
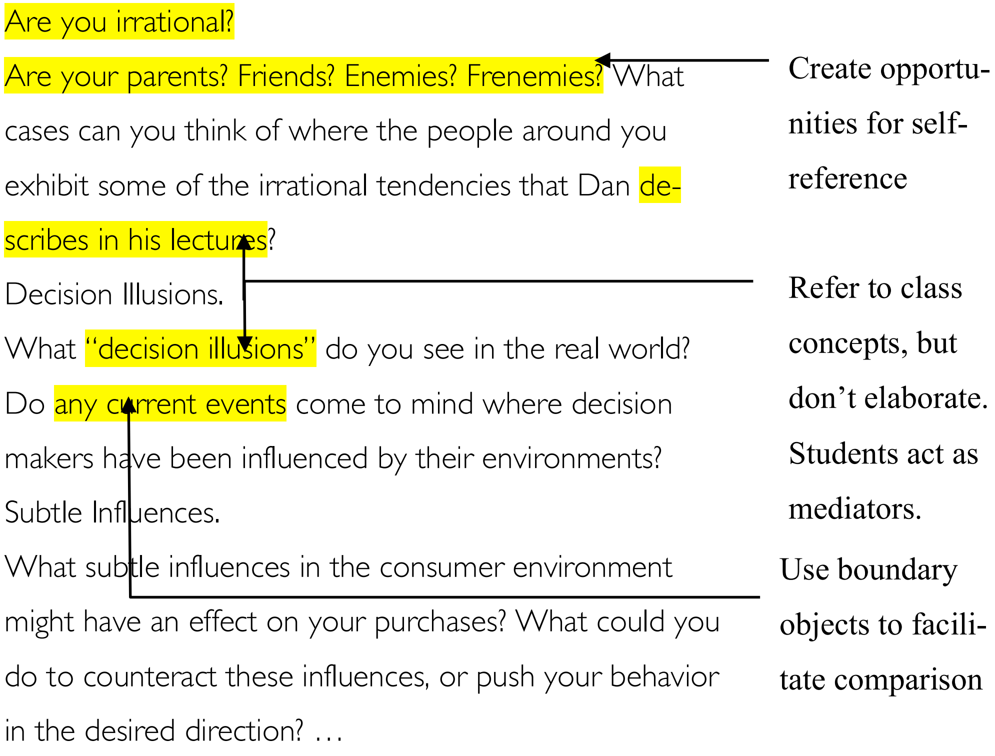


Figure 46: Across classes (a) Students from many countries participate in each six-person discussion (b) These students aren't just from neighboring countries, they are globally distributed.

scripts for discussion have a major impact on student engagement and learning [208]. Talkabout succeeds best when discussions create opportunities to highlight students' diverse experiences. Based on prior work, we developed three strategies to create discussion scripts or agendas, and refined them through deployments in seven massive classes. Figure 47 shows these strategies embodied in an excerpt from an *Irrational Behavior* agenda (the complete agenda is in Supplementary Materials). We discuss each strategy in turn.

Create opportunities for self-reference

Self-reference, when students actively relate class content to their own experiences and perspectives, increases concept elaboration, memory organization, and knowledge retention [54]. Talkabout agendas that employ self-reference ask students to share personal examples that embody class concepts. Self-reference is especially effective when students feel safe in discussing personal experiences. Talkabout groups are small by



Are you irrational?

Are your parents? Friends? Enemies? Frenemies? What cases can you think of where the people around you exhibit some of the irrational tendencies that Dan describes in his lectures?

Decision Illusions.

What "decision illusions" do you see in the real world? Do any current events come to mind where decision makers have been influenced by their environments?

Subtle Influences.

What subtle influences in the consumer environment might have an effect on your purchases? What could you do to counteract these influences, or push your behavior in the desired direction? …

Create opportunities for self-reference

Refer to class concepts, but don't elaborate. Students act as mediators.

Use boundary objects to facilitate comparison

Figure 47: Excerpt from discussion agenda in an Irrational Behavior discussion, showing examples of discussion-structuring strategies (highlighted)

design to encourage self-disclosure [209]. As each person shares with the group, it encourages peers to likewise disclose [210].

The globally distributed nature of discussions amplifies the benefits of sharing self-referential frames. After a discussion on prejudice in Social Psychology, one student wrote, “I think this may have been the first time the lady from Saudi Arabia had spoken to a Jew [referring to himself]”, showing her a different viewpoint. He added, “I told her about the prejudice from Christians I experienced growing up in [US state] in the 40's and the effect of segregation on blacks,” reflecting on his own experience. Students may see different self-referential frames with different groups. For instance, even though Social Psychology had only one Talkabout discussion (with multiple slots), 454 out of 2,553 participants in the Social Psychology class voluntarily attended multiple timeslots.

Highlight viewpoint differences using boundary objects

Talkabout prompts aim to make the differences between students’ perspectives salient. This encourages additional self-reference and re-evaluation of previously held theories, which in turn leads to deeper understanding [211].

To highlight differences, Talkabout discussion agendas call out boundary objects across geographical contexts. Boundary objects are objects or concepts that maintain their integrity across communities, and yet can be interpreted differently in different communities [212]. Everyday concepts, such as governments, companies/organizations or current events can serve as boundary objects. For instance, one student noted how discussing a ‘recent event’ yielded new perspective: “we were … joined by [a] Syrian. She provided…insight of the situation in Syria and how the media is exaggerating it… and how the society was quite liberal on Islamic practices (such as wearing the hijab).”

Leverage students as elaborators and mediators

When a prompt says less, students sometimes say more. Rather than reviewing every relevant concept, Talkabout discussion agendas reference concepts from class without any reminders of what they mean. These underspecified references lead students who have learned these concepts to elaborate, and to act as mediators with students who would have otherwise not understood them. This is similar to highly effective offline strategies like jigsaw classrooms, which also rely on peer-mediated learning and contact with dissimilar peers [29].

Creating opportunities for mediation also encourages students to ask about other class concepts they haven't understood. For instance, the Organizational Analysis class used "white flight" (a large-scale migration of white Americans to suburbs in the 1950s) as an example of an organizational problem faced by cities. In one Talkabout discussion session, we observed an American student translate the key ideas in this example to a European classmate by making an analogy to intra-European migration.

The anatomy of a Talkabout discussion

What is the nature of a Talkabout discussion session? We observed and recorded twelve Talkabout discussion sessions in *Organizational Analysis*. An abridged transcript from an Organizational Analysis class is in Supplementary Materials. Talkabout discussion sessions followed a pattern with clear roles and norms.

Discussions follow a distinct conversational pattern

Talkabout discussion sessions usually began with introductions. Since none of the participants knew each other, introductions were fairly formal and detailed. Participants typically shared their first name, their country of residence, and a brief description of their job. Because some participants arrived late to their session, this introduction phase was often repeated.

During these introductions, an informal moderator usually emerged (Refer to Appendix 1 for examples). Moderators often had experience with video-conferencing and a high-bandwidth connection. They exhibited leadership behaviors such as asking participants to introduce themselves, or even explicitly asking to moderate the conversation (e.g. "Shall I lead the conversation?")

After introductions, the informal moderator drew the group's attention to the instructor-provided discussion agenda. Even though agendas sometimes suggested a particular discussion order, participants did not follow it exactly. Instead, they would interpret the agenda for the major theme it embodied, and negotiate what they discussed first. Once students finished discussing a particular prompt, they returned to the agenda to decide the next topic.

While Talkabout discussion sessions were designed to last 30 minutes, the median length of the discussion was 58 minutes (Figure 48). With these longer discussions, students discussed topics that were marked optional, or chose to discuss two topics when the agenda asked only one etc. Many groups also spoke about the class in general after the assigned topics. Conversations typically ended soon after the informal moderator (or a talkative speaker) left the discussion, or when no one in the group suggested a topic to discuss next. As they left, participants often shared how they enjoyed talking to the group, or taking the class. Moderators sometimes encouraged the group to stay in touch after the discussion (e.g. "With the other hangouts, we all added each other on LinkedIn… I've already added [name]. If you'd like, feel free to add me.")

Speakers and spectators

Students seemed to decide early on whether they primarily wanted to speak during the discussion ("speakers"), or listen to the discussion ("spectators"). Spectators often signaled their intent by muting their microphones (this showed a "microphone muted"

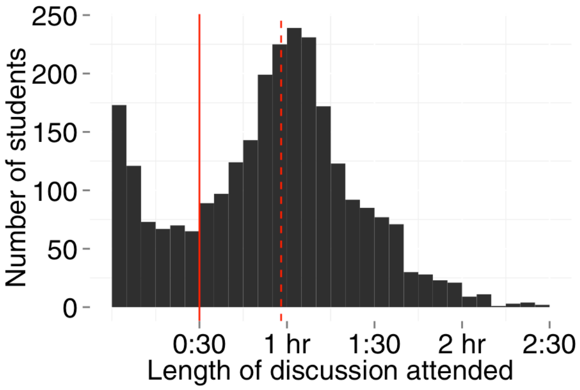


Figure 48: Across classes, students participated in discussions much longer than instructions indicated. The solid red line is the recommended duration for discussion (30 min), the dashed line is the median discussion time (58 min).

icon to others in the discussion).

Speakers tended to be native English speakers or have faster Internet connections. Their discussion was conversational, with overlapping turns similar to face-to-face conversation. Spectators spoke less frequently with longer non-overlapping turns, but were not passive participants. When spectators had trouble finding the right words (e.g., if they were non-native speakers), speakers often suggested words, or encouraged them to continue.

Participants with low-bandwidth connections generally assumed the spectator role and often used the text chat feature in the Google Hangout to "speak" in the discussion. Speakers (usually the moderator) would notice the text, and speak it aloud to the other participants. Both speakers and spectators used text-chat to demonstrate active listening without interrupting the speaker via audio (for example, a student wrote, "Working in [company] must be really cool. Thanks for sharing :)").

A shared video channel forces a single conversation. Still, students sometimes used text-chat as a way for non-discussion related talk, such as exchanging contact information or LinkedIn profiles.

Study 1: Do discussions help performance?

It is not obvious that the benefits of peer discussions [213], [214] would transfer to an online environment. In these environments, peers have vastly different backgrounds and no prior interaction with each other. Therefore, our first study measures the benefits of participation in online discussions. Later experiments measure how these benefits vary with geographic diversity in discussion groups.

With many educational practices, it is difficult to draw a causal link between participation and student learning. For instance, students may self-select to participate. To combat this bias, we use a control condition in which interested students are actively prevented from discussing. Furthermore, we use an *intention-to-treat* analysis that recognizes that some students will not participate, even when given the opportunity. Therefore, this analysis asks: after controlling for students that don't discuss given an opportunity, are discussions effective? Such analysis is common in clinical trials, where patients that are randomly assigned to a treatment group are included in the analysis even if they do not take their medication. Because intention-to-treat analyses take non-compliance into account, they result in conservative estimates of a drug's effectiveness.

Method: wait-list control

In a between-subjects experiment, we randomly assigned students in the Organizational Analysis class on Coursera to either a *Discussion* condition, or to a *Wait-list* condition. This assignment occurred when they signed up for a discussion time on Talkabout, after consenting to participate in the study.

Students in the Discussion condition were allowed to participate in discussions starting in Week 1, while those on the wait-list were not allowed to participate in discussions until Week 5. This setup results in two discussion opportunities (Week 1 and Week 3) where a subset of students was prevented from participating. Even though some participants in the Discussion condition did not attend discussion, they were included in the intention-to-treat analysis.

Hypotheses and measures

We hypothesized that participating in a Talkabout discussion session would motivate students to engage with other course components. Prior work similarly finds that discussions motivate students to engage with in-person classes [214]. To measure engagement, we check whether the student participated in the course quiz due the day after discussion. Recall that participation in MOOCs is entirely voluntary, and several classes have battled with attrition [215]. Quizzes are a high-effort activity that most MOOC learners don't participate in: only 22.8% of students who watched a lecture video also participated in a quiz. This makes quizzes suitable as a high-effort engagement measure [205], [216].

We further hypothesized that students in the Discussion condition would do better on the quiz, aided by the self-reference, reflection and revision of class concepts.

Participants

Overall, 1,002 students were assigned to the Discussion condition, and 122 to the Wait-list condition. We used an unbalanced design to maximize the number of students who benefited from discussions. Of those in the discussion condition, 397 attended a discussion.

Results: Discussion increases class participation, marginally improves grades

Students in the Discussion condition were more likely to take the quiz. A logistic regression indicated that odds of taking the quiz were 1.46 times higher for the Discussion condition (Wald z*=1.97, *p<0.05*). Students in the Discussion condition also did marginally better on the quiz (t(1122) = 1.89, *p=0.06*)[8]. The average improvement was 16.7%.

Thus, even accounting for students who do not follow through, discussions help students stay engaged in the course and perform better on related assessments.
While Talkabout participation improves engagement, this effect seems short-lived. Students who participate in a Talkabout one week are not more likely to participate in the quiz the following week: Wald z*=1.61, *p=0.10*. We also found no significant improvement in quiz scores for the quiz due the following week.

Would participating in multiple Talkabout discussion sessions improve these short-term benefits? As is typical with online classes, many students shopped the first weeks, and only 113 students in the discussion condition attended the second discussion (397 attended the first week). Therefore, our intention-to-treat analysis lacks the statistical power to capture any benefits of participating in multiple discussions. Also, while the wait-list design can control for intent to participate, students who actually participate in discussions may still differ from those who don't (e.g. they could be

[8] While only marginally significant (p<0.10), we include this result because it is suggests opportunities for future work.

more motivated). An intention-to-treat analysis estimates effects by assuming participants' distribution (e.g., for motivation) are similar in the wait-list and treatment groups due to randomized assignment, but this experiment does not verify this assumption.

The results of this study suggest that performance on class quizzes may improve even with limited participation, and that discussions improve student engagement. Do these effects depend on the participants in the discussion? Given our hypothesis that geographic diversity should help learning, our next study investigates the effect of discussants' geographic diversity on course performance.

Study 2: Does diversity help performance?

Study 1 established that participating in Talkabout discussions improves class engagement. Is geographic diversity causing this effect? In a second, between-subjects experiment, Talkabout's group-assignment algorithm randomly assigned students either to a single-region group or a multi-region group. Participants regions were determined by the five World Bank regions, as well as three regions to capture North America, Western Europe and the South Pacific. The *Same-region* condition grouped students with others from their region. The *Multi-region* condition grouped students from anywhere in the world. We discarded data from the South Pacific region because it had few participants.

Participants and setup

55 students in the Organizational Analysis class participated. When students logged on to the site, we recorded their IP address, found their location based on IP, and randomly assigned them to the Multi-region high-diversity or the Same-region low-diversity condition. Students were then grouped into discussion groups with a maximum of six participants.

Measures

To measure conceptual understanding, we invited students to fill out a questionnaire immediately after the discussion; 43 participated. We asked students to answer to the best of their ability, but informed them that their answer would not affect their course grade. This survey had one open-ended question which required critical thinking and an understanding of concepts discussed in the session ("Where would you want to position yourself if you wanted leverage over the flow of "information" in a social network—centrally, peripherally, or in a bridging position. Why?"). We scored this question in consultation with the teaching assistant of the course. The average score was 47% (combining both conditions). We use students' grade in a prior class quiz as a measure of prior performance (we ignore data from one participant, who did not complete the quiz). The questionnaire also asked questions about how much they liked their discussion, and how much they felt they learned from it.

Hypothesis

Students in the Multi-Region, high-diversity, condition were exposed to more contrasting viewpoints and self-reference than discussions in low-diversity groups. Thus, we hypothesized that members of more geographically diverse groups would have higher scores on the post-questionnaire.

Manipulation check

The median number of countries in the same-region condition was two (both from the same geographical region), while the median in the multiple-region condition was 4. Does large geographical distance imply a diverse group? Some World Bank regions are large, so we examined if multi-region groups had more differing national viewpoints than same-region groups, taking into account how economic opportunities and educational experience influence everyday experience [217], as do cultural values [198].

We used each participant's country to map them onto diversity attributes used in cultural psychology and political science. We use countries as our unit of analysis because they have a consistent typology of collectivistic or individualistic culture [189], organizational attitudes such as inter-personal dependence and criteria for fulfillment [218], economic development [190] and life expectancy [219]. While each country is diverse, within-country differences are smaller than between-country differences [189], making this by-country analysis feasible.

We compared countries of participating students on three dimensions: cultural values, income, and pupil-teacher ratios in primary school. As a measure of cultural values, we used the mean overall secular values for each country from the World Values Survey [220]. Countries with lower scores have societies that emphasize religion, traditional family values, and collectivistic thinking. The average pair-wise difference between participants' countries on the overall secular values scale was lower in the same-region condition than in the multi-region condition, Wilcoxon W=407.5, $p<0.05$ (*same-region* mean: 0.022, equivalent to the difference between the US and Romania, *multi-region* mean: 0.031, equivalent difference: US and Thailand).

Students' countries in the *Multi-region* condition had marginally higher differences in income levels compared to those in the *Same-region* condition (t(74)=1.81, $p=0.07$; log-transformed because income distribution is log-normal [221]). Using data from the World Bank [207], the median per-capita annual income differed on average by $8,120 (PPP) in the *same-region* condition, approximately the difference between the US and Canada. The average difference in the *multi-region* condition was $20,495 (PPP), approximately the difference between the US and Israel.

Lastly, students' countries in the multi-region condition had greater pairwise variation in educational experience, as reflected in primary school pupil-teacher ratios (t(74)=2.00, $p<0.05$). Using World Bank data [207], the median differences in the pupil-teacher ratios in the same-region condition were 2.91 (approximately the difference between schools in the US and Canada), while the median difference in the multi-group condition was 5.91 (the difference in schools between the US and Russia). Collectively, these analyses suggest that multi-region groups brought more diverse experiences and backgrounds to their discussions.

Results: Students in diverse groups perform better

On a 7-point Likert scale question, students in the *high-diversity* condition rated their discussion as more enjoyable than those in *low-diversity* (Mann-Whitney U=140.5, $p < 0.05$). They also reported learning marginally more from their discussion partners on a different 7-point Likert scale (Mann-Whitney U=160.5, $p = 0.08$).

Based on the grades in the post-quiz, an ordinary-least-squares linear model showed that after controlling for prior performance, students in the *high-diversity* condition out-performed those in the *low-diversity* condition, (β=0.41, F(1,37)=2.31, $p<0.05$, adjusted R^2=0.11). A post-hoc comparison also found that students in discussions with more countries did better in both conditions. Using an ordinary-least-squares linear model, we found that the number of countries in the discussion was predictive of the quiz score (β=0.15, F(1,36) = 2.57, $p<0.01$, adjusted R^2= 0.14).

This result suggests that even countries in the same geographical region add meaningful diversity. This may be because regions are too large and diverse (e.g. the *Latin America and Caribbean* region has 35 countries). Therefore, counting countries rather than regions may provide a better measure of diversity.

However, this experiment only measures the immediate effects of diversity in a single class. Do geographically diverse discussions have a longer-term effect, and do these benefits generalize across classes? We now describe a longitudinal deployment that evaluates the effect of diverse discussions on grades in actual course tests over periods of weeks.

Study 3: Large scale field experiment

In Study 3, we sought to confirm and expand upon Study 2’s diversity effect across more classes and with more students. In doing so, we trade off some of Study 2’s experimental control in exchange for a much larger sample. We conducted our experiment across two large online classes, Organizational Analysis and Social Psychology.

Participants

In the Social Psychology class, 2,025 students participated. In the Organizational Analysis, 397 students participated.

All students in the Organizational Analysis class who wanted to participate in discussions used Talkabout. By the instructor's request, the Social Psychology class also allowed students to choose an in-person discussion instead. In-person discussants received the same discussion agenda and directions as online discussants. 2,037 students reported participating in an in-person discussion. Except for qualitative comparisons between online and in-person discussions, we ignore their data. It is possible that online discussions attracted students who believed they would benefit more from a diverse discussion. However, the main results of this study were consistent across both classes.

Method

Similar to Study 2, Talkabout grouped students into discussions. However, students were not explicitly grouped into high- and low-diversity conditions. Instead, this study used a simpler approach where Talkabout collected participants in order of arrival. When a group had six students, Talk-about launched a new group. This setup assigns participants to diversity levels in a random fashion. Participants in both classes had no control over who their discussion partners were, and therefore had no control over the level of geographic diversity in their discussion.

The two classes implemented different schedules for their discussions. Social Psychology held discussions for one week at the end of class, two weeks before the final exam. Organizational Analysis had discussions throughout the class, starting from the first week. This variety allows us to understand the effect of Talkabout both for highly motivated students who remain active at the end of class, and for enthusiastic, but potentially uncommitted learners.

Hypotheses and Measures

We hypothesized that participating in more geographically diverse Talkabout discussions would lead to better course performance, as students became more active thinkers through conversations with diverse students. In addition, given our results in Study 1, we hypothesized that students in more diverse discussions early in the class would stay engaged with the class for longer.

To measure *geographic diversity*, we use the number of countries in a discussion as a coarse but useful metric. While students using Talkabout may be systematically different from the median resident of their country (they can afford an Internet connection), national cultures still importantly shape their thoughts and actions [222].

To measure *performance*, in Social Psychology, we used the final exam score. The final exam was a 50 multiple-choice question test (see Appendix 2 for a sample of questions). The instructor created this exam independently with no input from the research team. The Organizational Analysis class had weekly quizzes due every Sunday, which we use as a performance measure. The instructor independently created these quizzes in a previous run of the class (before Talkabout was designed), and they were used unchanged in the experimental class. The first Talkabout session was one day before the first quiz was due. We analyze the first two weeks' quizzes. The first quiz had 19 multiple-choice questions; the second had 16 (see Appendix 2). Finally, both classes invited students to participate in a post-discussion survey about their experience.

Analysis procedure

For both classes, we built an ordinary-least-square linear regression for performance based on the number of countries in the discussion. Because the number of discussants and number of countries is collinear (R^2=0.81 and 0.88 in the two classes), we only analyzed groups of six students. We controlled for each student's prior performance in class if any previous quizzes had occurred. Our model for the first week's quiz in Organizational Analysis had no measure for prior performance (model R^2=0.003). The model for the second quiz (R^2 = 0.11) used the score in the first quiz as a prior-performance metric. The model for the Social Psychology class (R^2=0.05) used a student's total grade in all assignments before the final exam as a prior-performance metric.

Results

Our analysis finds support for the first hypothesis: students perform better on tests after a more geographically diverse discussion. We find no support for our second hypothesis that diverse discussion improves retention in the long-term.

High-diversity discussions improve scores

In both classes, more diverse discussions led to higher exam grades (Table 10). In Social Psychology, on the final exam out of 50 points, each additional country adds an approximate β=1.78 points (2.4% of the final grade) to a student's final exam score (t(129)=1.78, *p=0.01*). In Organizational Analysis, on the Week 2 quiz out of 16 points, each additional country yields β=0.39 points (3.6%) to the quiz score (t(110)=2.03, *p<0.05*). However, from the model for the Week 1 quiz (without a prior-

	β	F	p-value
Organizational Analysis: Week 1 Quiz (R^2 = 0.003)			
Intercept	15.7	21.51	<0.001
Number of Countries	0.11	0.76	0.46
Organizational Analysis: Week 2 Quiz (R^2 = 0.11)			
Intercept	8.11	4.33	<0.001
Week 1 grade (z-scored)	0.78	2.81	< 0.001
Number of countries	0.39	2.03	0.02
Social Psychology: Final Examination (R^2 = 0.05)			
Intercept	27.20	7.00	<0.001
Pre-final grade (z-scored)	0.91	1.30	0.19
Number of countries	1.78	2.34	0.01

Table 10: After controlling for prior performance, more countries in a discussion lead to better grades, in both Social Psychology and Organizational Analysis.

performance measure), we do not see any significant effect of diversity on score. Prior performance helps capture sufficient variation to make diversity statistically distinguishable from a null hypothesis.

Benefits of diverse discussions last roughly two weeks

In the Organizational Analysis class, while geographic diversity leads to better quiz scores one week after discussion (Week 2 quiz), we did not find any significant effects into Week 3. Similarly, we built an ordinary-least-squares linear model for predicting how many weekly quizzes a student would participate in, based on the number of countries in their first discussion. We found no significant effect ($t(130) = -0.49$, $R^2 < 0.001$). Similar to results from Study 2, this suggests that the benefits of a diverse discussion only persist for a short duration.

Geographic diversity leads to new perspectives

Post-discussion, a survey asked participants about the best part of their discussion. Two independent raters coded 100 responses about whether comments mentioned participant diversity: 51% mentioned it (Cohen's κ =0.7, z=7.04, $p< 0.001$). Students noted that diversity yielded different experiences and examples and perspectives, which challenged ones held by students. A Social Psychology student wrote how they learned that "…in China it is a custom for married women to keep their surnames, thus I [now] think women changing their surnames when married in other countries has something to do with sexism." An Organizational Analysis student said, "It was interesting to hear about organizations in Australia, Ukraine, Israel, Indonesia, and Canada. Similar issues appear everywhere regarding decision-making"

Gender representation does not influence scores

In prior work, the proportion of females participants affected collaborative group outcomes [223]. However, in our study, female participation did not affect performance after controlling for the number of countries in each group. Adding the proportion of female participants to the Organizational Analysis class model for the Week 2 quiz did not improve model fit, and the effect of gender was not significant: t(100)=1.1, *p=0.26*. The Social Psychology class shows a similar non-significant effect: t(128)=0.62, *p=0.53*.

Other non-significant factors

We test the following variables in isolation; all were non-significant with *p>0.50*. We found no significant effect of the arrival order of participants on either the diversity in their group, or the benefits of diversity on course grades. We also find no evidence that diverse discussions had larger benefits for either gender. Finally, there was no significant correlation between how early students signed up for a discussion and their benefits.

Other measures of geographic diversity

The results of our analysis were consistent when we used other measures such as the pairwise distance between participants' locations. We use the number of countries while describing results because it is more interpretable.

Limitations

This experiment included two classes, Social Psychology and Organizational Analysis. Both classes used Talkabout in discussions focused on critical thinking and sense-making. As such, evidence that geographically diverse discussions improve engagement and learning may not generalize to classes that emphasize procedural knowledge (e.g. Corporate Finance), or classes where benefits from global perspectives are smaller (e.g. physics). That said, even the most procedural topics require critical thinking and judgment, and as many instructors have found, topics like physics that seemingly don't benefit from global perspectives may still benefit from discussions [224], [225]. Geographic diversity encodes many other kinds of diversity, e.g., economic opportunities, cultural values, and education experience. Each of these dimensions may have differing benefits for online classes. Future work could build theory that differences matter when.

Study 1: Discussion participation with a wait-list control	**Participating in a video discussion with peers increases participation in quizzes and marginally improves performance.**
Study 2: Controlled manipulation of geographic diversity	**Students in high geographic diversity discussion groups perform higher.**
Study 3: Large-scale study of geographic diversity	**High geographic diversity discussions lead to improved short-term performance in two classes, but do not improve multi-week retention.**

Table 11: Summary of experimental results

Discussion

It can be difficult to demonstrate measurable learning effects using design interventions in online courses. For example, while it is possible to increase student involvement in forums [226], improving grades and retention has remained challenging [205], [216]. However, Talkabout increases both learning and engagement (Table 11). One reason for this improvement may be that Talkabout developed a pedagogical approach alongside the software. In pilots without meaningfully structured discussions, it fared poorly. Furthermore, Talkabout builds a social environment and an opportunity for reflection. It does this via a medium that is known to build trust [200] and is suited for open-ended discussions [201], such as those leading to sense-making [204].

Geographic diversity's direct effect is in students meeting people from other world regions. It is associated with changes in several other diversity measures (e.g., cultural values, economic opportunity, and educational experience). This chapter demonstrates that geographic diversity indeed impacts these other measures. However, there may be other causal pathways involved. It is possible that students who differ in geographical location still have similar socio-economic backgrounds, and students who live very close may be very different. Future work can develop more nuanced diverse experiences.

Talkabout also points to the benefits of using video for geographically diverse discussions. Video conferencing creates a middle ground of immersion in another culture. With complete immersion in an in-person setting, the norms and views of the majority are pervasive [227], [228]. Students with a minority viewpoint in a fully-immersive experience may find themselves confronted with the choice to either embrace the majority culture (suppressing their own), or reject it and flounder [229]. On the other hand, with the minimal immersion, say, of lectures, students may ignore alternate viewpoints as a mere academic exercise. Video-conferencing may occupy an attractive middle ground: it is interactive, compelling students to engage with their diverse classmates and reflect upon their contact [192]. One student told us in an interview, "Talkabout helps bring the class together -- it makes the learning tangible and real…you are interacting with other people, who are experiencing a lot of different things."

Video-based discussions are not without their problems today. Some countries (*e.g.,* Iran) restrict access to Google Hangouts, low-bandwidth connections degrade the student experience, and installing video-conferencing software remains challenging for some students. However, these technological limitations are likely to lessen as bandwidth becomes more plentiful and software comes pre-installed.

Comparing in-person and online discussions

Recall that Social Psychology allowed students to choose to run their discussion in person instead of online. Students participating in the in-person discussions often turned to close friends and relatives. The shared context made the conversation friendlier. For instance, one participant remarked, "I really like the discussion because it was with my friends… It was really easy to start the discussion." In-person discussions also had lower geographic diversity. One student summarized, "Being from the same age group, social level and from the same community; we had very much similar views about the topics in hand." Students reported difficulties scheduling discussions and keeping them on-topic. One remarked, "We had to reschedule a couple of times [before we could meet].” And with friends, “Turning a conversation towards a scientific discipline such as social psychology was hard and a bit artificial…” Another remarked, “Members were my family... and speaking about some things is not easy!”

The design space of online peer conversations

Talkabout currently implements a particular design for online discussions. To arrive at this design, we explored a number of different decisions in this design space (Table 12).

Always-available discussions lack critical mass

Always-available and unscheduled discussions in classes may enable students to talk with a remote partner whenever they have a question or thought. To test the feasibility of this idea, we created a version of Talkabout where students could sign up for an immediate discussion. If another student indicated their availability within the next hour, Talkabout would email both to set up a discussion.

We tested this version in the Think Again philosophy and argumentation class over a three-day period. Of the 2,940 who saw the opportunity, 54 students signed up. Unfortunately, only 5 students overlapped within the one-hour window. This suggests that MOOCs attract many students, but their presence on the course site does not spontaneously overlap. Therefore, Talkabout instead adopts a *bus stop model* where discussions occur at regular time intervals, making critical mass more likely.

Students prefer to negotiate roles informally

Prior work suggests including a designated discussion facilitator to attend to group dynamics in distributed discussions [230]. Could formal facilitators improve Talkabout discussions? We conducted a between-subjects experiment with two conditions (n= 80) in the Organizational Analysis class. In the *facilitator* condition, all participants in a Hangout saw a button to volunteer to be a discussion facilitator. When a student volunteered, the system would show them facilitation tips. Other participants

Design Dimension	**Choices**		
Same discussants every time?	Yes	No	When possible
Group size	Small		Large
Discussion guidance	None	Guidelines and prompts	Scripts
Role negotiation	Instructor specifies	Technologically mediated	Informal
Discussion scheduling	On-demand any time	Bus stop: regular intervals	Same time every week

Table 12: Talkabout's current implementation highlighted in dark blue, design choices that we found to be worse are highlighted in (light) red.

saw a message that the volunteer was facilitating the discussion. In the *control* condition, students were not shown the button to volunteer. Of the 40 students in the *facilitator* condition, seven volunteered. An intention-to-treat analysis showed a trend toward students in the *facilitation* condition feeling the discussion was less motivating (Mann-Whitney U = 191.5, *p=0.09*), and a trend toward less willingness to meet the same group again (U=191.5, *p=0.09*). These results suggest that fluid negotiation of moderation may work better than a formal facilitation role.

Rigidly enforced scripts lower satisfaction

Prior work in CSCL suggests that structuring collaboration between students using instructions or scripts yields improved learning [208]. What is the right degree of scripting? In a between-subjects experiment (n=82) in the Organizational Analysis class, we explored the benefit of an enforced script. In this condition, Talkabout only showed the current discussion topic, and participants needed to click a button to indicate completion and advance to the next topic. The *control* condition agenda showed all topics at once.

Of the 50 students in the *enforced-script* condition, only 4 clicked the "next topic" button even once. In the post-survey, students also reported they felt the discussion was less motivating (Mann-Whitney U= 193, p=0.07), and that they were less willing to meet the same group again (U = 191.5, p= 0.08). This suggests that enforcing a discussion order may undermine the social benefits of Talkabout [231].

Same-partner discussions have inadequate participation

In the in-person classroom, it is common practice to assign students to groups with fixed membership for the duration of a project or series of discussions throughout a course [214]. Repeated interactions in such groups build trust and rapport [232]. By contrast, non-persistent groups lack familiarity but expose students to different viewpoints.

In a between-subjects experiment in the Think Again class (n=522), we randomly assigned students to either a *persistent* or *control* condition. The persistent condition assigned students to the same group for every discussion. The control condition assigned them to a group when they arrived to the site, as described previously. Students in both conditions attended the same number of discussions (μ=0.46, t(522)=0.33, *p=0.73*). However, as students dropped the class, the size of discussion groups in the persistent condition kept shrinking until they were no longer viable. While 27% of the control groups had at least 5 discussants, only 2% of the persistent groups did (t(81)=4.67, *p<0.001*). Therefore, our discussion strategies structure discussions to leverage changing partners. The next step might be to forge a middle ground where Talkabout prefers familiar partners but adapts groups if previous partners drop out.

Conclusion

This chapter suggests that the geographic diversity in online classes can be an educational asset. Instead of becoming a handicap, distance can expose students to others and to other ways of thinking. However, leveraging the diversity of online environments requires careful design. This chapter describes one such approach, Talkabout, which uses video chat to create discussions between students across the world. Embracing and designing for diversity can enable other innovations. For instance, instructors could leverage students as co-creators and draw on students' local observations to showcase how course concepts arise differently around the world. Likewise, international relations or security courses might launch a global crisis simulation with each student representing their own region. These educational experiences offer a glimpse of the potential of thinking “beyond being there” [42]. They are not just leveraging geographic diversity—they would be impossible without it.

Chapter 7
Adoption challenges for global peer learning systems and how to address them[9]

Many online classes use video lectures and individual student exercises to instruct and assess students. While vast numbers of students log on to these classes individually, many of the educationally valuable social interactions of brick-and-mortar classes are lost: online learners are “alone together” [233]. This dissertation demonstrates how to introduce social interactions into MOOCs.

Social interactions amongst peers improves conceptual understanding and engagement, in turn increasing course performance and completion rates [11, 20, 22, 26, 28]. Benefits aren’t limited to the present: when peers construct knowledge together, they acquire critical-thinking skills crucial for life after school [237]. Common social learning strategies include discussing course materials, asking each other questions, and reviewing each other’s work [238].

So far in this dissertation, we have discussed how to design systems for peer interactions that help students who use them learn better. But given that participation in MOOCs is largely voluntary, how can we encourage students to use these systems in the first place?

[9] A version of this chapter was originally published as an article in the proceeding of the ACM Conference on Learning at Scale, 2015 as [292].

Three impediments to adoption… and remedies

Educational peer platforms connect students in massive online classes in order to discuss course topics, reflect on others' ideas, and build esprit de corps [5, 23]. Over the last two years, three challenges have consistently recurred as we have introduced peer learning into massive online classes.

First, many courses falsely assume that students will naturally populate the peer learning systems in their classes: "build it and they will come". This assumption often seems natural; after all, students naturally engage with social networks such as Facebook and Twitter. However, students don't yet know why or how they should take advantage of peer learning opportunities. Peer learning platforms sit not in a social setting, but in an educational setting, where participation is often effortful, and it may take years for the benefits to become apparent.

Because benefits of participation are not immediately apparent, educational settings has a different incentive structure than a socialization setting. For instance, many American college graduates retrospectively credit their dorms as having played a key role in their social development [240]. Yet, universities often have to require that freshmen live in the dorms to ensure the joint experience. A similar reinforcing approach may be useful online, integrating peer-learning systems into the core curriculum and making them a required or extra-credit granting part of the course, rather than optional "hang-out" rooms.

The second challenge is that students in online classes lack the ambient social encouragement that brick-and-mortar settings provide [241]. The physical and social configurations of in-person schools offer many opportunities for social encouragement (especially to residential students) [11, 22]. For example, during finals week, everyone else is studying too. However, other students' activity is typically invisible online, so students do not form the descriptive norms, or benefit from the encouragement of seeing others attend classes and study [13, 15]. We hypothesize that in the minimal social context online, software and courses must work especially hard to keep students engaged through highlighting co-dependence and strengthening positive norms.

The third challenge is that instructors can, at best, observe peer interactions through a cloudy telescope: summary statistics offer few visible signals beyond engagement (e.g. course forum posts and dashboards) and demographics. Student information is limited online [243], and knowing how to leverage what demographics instructors do know is non-obvious. In-person, instructors use a lot of information about people to structure interactions [244]. For example, instructors can observe and adapt to student reactions while facilitating peer interactions. The lack of information in online classes creates both pedagogical and design challenges [245]. For instance, in an online discussion, do students completely ignore the course-related discussion prompts and, instead, talk about current events or pop culture? To address such questions, teachers must have the tools to enable them to learn how to scaffold peer interactions from behind their computers.

This chapter addresses these three logistical and pedagogical challenges to global-scale peer learning (Figure 49). We suggest socio-technical remedies that draw on our experience with two social learning platforms – Talkabout and PeerStudio – and with our experience using peer learning in the classroom.

We report on these challenges with both quantitative and qualitative data. Quantitative measures of efficacy include sign-up and follow-through rates, course participation and activity, and participation structure and duration. Qualitative data includes students' and instructors' comments in surveys and interviews. We describe how peer learning behavior varies with changing student practices, teacher practices, and course materials.

How students use two peer learning platforms

Below, we provide an overview of the two peer systems that form the basis of this chapter. The first, Talkabout (Chapter 6), brings students in MOOCs together to discuss course materials in small groups of four to six students over Google Hangouts [239]. Currently, over 4,500 students from 134 countries have used Talkabout in 18 different online classes through the Coursera and Open edX platforms. Students join a discussion timeslot based on their availability, and upon arriving to the discussion, are placed in a discussion group; on average there are four countries represented per discussion group. We have seen that students in discussions with peers from diverse regions outperformed students in discussions with more homogenous peers, in terms of retention and exam score [234]. We hypothesize that geographically diverse discussions catalyze more active thinking and reflection.

The second platform, PeerStudio (Chapter 5), provides fast feedback on in-progress open-ended work, such as essays [37]. Over 4,000 students in two courses on Coursera and OpenEdX have used PeerStudio. Students submit a draft, an essay for example, and are then prompted to review two other drafts. After completing two reviews, they can access the feedback on their essay. With PeerStudio, students can receive formative feedback on their draft work within hours. A randomized controlled experiment showed students created better revisions when they have rapid feedback from their

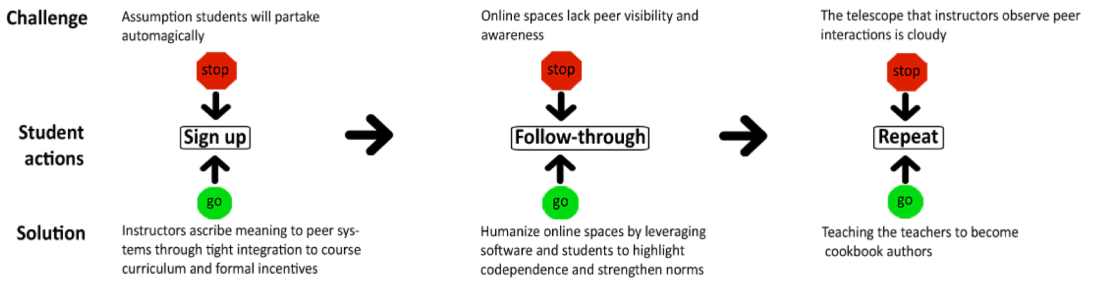

Challenge
Assumption students will partake automagically
Online spaces lack peer visibility and awareness
The telescope that instructors observe peer interactions is cloudy
stop
stop
stop
Student actions
Sign up
Follow-through
Repeat
go
go
go
Solution
Instructors ascribe meaning to peer systems through tight integration to course curriculum and formal incentives
Humanize online spaces by leveraging software and students to highlight codependence and strengthen norms
Teaching the teachers to become cookbook authors

Figure 49: The challenges and remedies of adoption of peer learning systems presented in this chapter.

peers, on average 20 minutes in our deployments at scale.

Social capabilities do not guarantee social use

Peer learning systems share many attributes with collaborative software more generally [246]. However, the additional features of the educational setting change users' calculus. Throughout the deployments of our platforms, we've observed different approaches that instructors take when using our peer systems in their classes.

Often, instructors dropped a platform into their class, then left it alone and assumed that students would populate it. For example, one course using Talkabout only mentioned it once in course announcements. Across four weeks, the sign-up rate was just 0.4%, compared to a more successful sign-up rate of 6.6% in another course; sign-up rate being the number of students who signed up to participate in the peer system out of the number of active students (students who watched a lecture video) in the course. Low percentages represent conservative estimates as the denominator represents students with minimal activity. When this theme recurred in other Talkabout courses, it was accompanied with the same outcome: social interactions languished. Why would instructors who put in significant effort developing discussion prompts introduce a peer learning system, but immediately abandon it?

Interviews with instructors suggested that they assumed that a peer system would behave like an already-popular social networking service like Facebook where people come en masse at their own will. This point of view resonates with a common assumption that MOOC students are extremely self-motivated, and that such motivation shapes their behavior [125], [247]. The assumption seemed to be that building a social space will cause students to just populate it and learn from each other.

However, peer-learning systems may need more active integration. The value of educational experiences is not immediately apparent to students, and those that are worthwhile need to be signaled as important in order to achieve adoption.

Chat rooms underscored a similar point of the importance of pedagogical integration. Early chat room implementations were easily accessible (embedded in-page near video lectures) but had little pedagogical scaffolding [248]. Later, more successful variants that strongly enforced a pedagogical structure were better received [249].

Peer software as learning spaces

Even the best-designed peer learning activities have little value unless students overcome initial reluctance to use them. Course credit helps students to commit, and those who have committed, to participate. Consider follow-though rates for Talkabout: the fraction of students who attend the discussion out of the students signed up for it. In an international women's rights course, before extra credit was offered, Talkabout follow-through rate was 31%. After offering extra credit, follow-through rate increased to 52%. Combined with other strategies from this chapter, we've seen formal incentives raise follow-through rates up to 64% in other classes.

Faculty can signal to students what matters by using scarce resources like grade composition and announcements. We hypothesize that these signals of academic importance and meaning increase student usage. For example, in a course where the instructors just repeatedly announced Talkabout in the beginning, 6.6% of active students signed up, a large increase from the 0.4% sign-up rate when there was only one mention of Talkabout.

We saw similar effects with PeerStudio. When participation comprises even a small fraction of a student's grade, usage increases substantially. In one class where PeerStudio was optional, the sign-up rate was 0.8%. The fraction of users was six times higher in another class where use of PeerStudio contributed to their grade: the

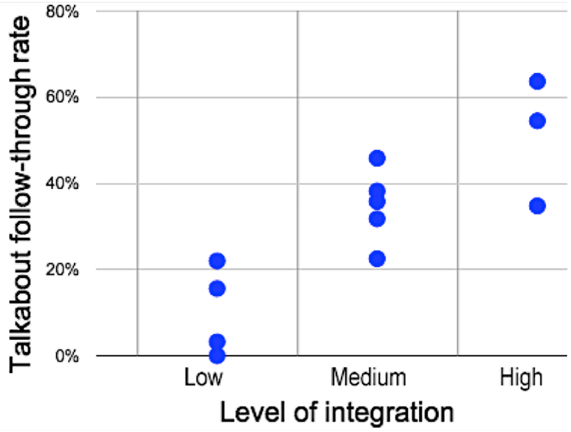


Figure 50: Follow-through rate from 12 Talkabout courses increases as integration increases.

sign-up rate was 4.9%. To maintain consistency with insights from Talkabout, sign-up rates for PeerStudio also represent the number of students who consented to participate in peer feedback activities out of the number of active students (students in the course who watched a lecture video).

Students look up to their instructors, creating a unique opportunity to get and keep students involved. One indicator of student interest is if they visited the Talkabout website. Figure 4 shows Talkabout page views after instructors posted on the course site discussing Talkabout, and a decrease in page views when no announcement is made. Talkabout traffic was dwindling towards the end of the course, so the instructor decided to offer extra credit for the last round of Talkabout discussions. During the extra-credit granting Talkabout discussions, page views increase around twofold the previous four rounds.

To understand how pedagogical integration and incentives, and follow-through rate interact, we divided 12 Talkabout courses into three categories, based on how well Talkabout was incentivized and integrated pedagogically (see Figure 50). Courses that never mentioned Talkabout or mentioned it only at the start of the course are labeled “Low integration”. Such courses considered Talkabout a primarily social opportunity, similar to a Facebook group. Few students signed up, and even fewer actually participated: the average follow-through rate was 10%. The next category, “Medium integration,” was well integrated but poorly incentivized, classes. These classes referred to Talkabout frequently in announcements, encouraged students to participate, and had well-structured discussion prompts, but they had no formal incentive. Such classes had an average follow-through rate of 35%. Well-incentivized and integrated classes, “High integration,” offered course extra credit for participation and continuously discussed Talkabout in course announcements, and averaged 50% follow-through rate.

This visualization highlights the pattern that the more integrated the peer learning platform is, the higher the follow-through rate is. We have found that offering even minimal course credit powerfully spurs initial participation, and that many interventions neglect to do this. As one student noted in a post-discussion survey, “I probably wouldn't have done it [a Talkabout session] were it not for the 5 extra credit points but I found it very interesting and glad I did do it!”

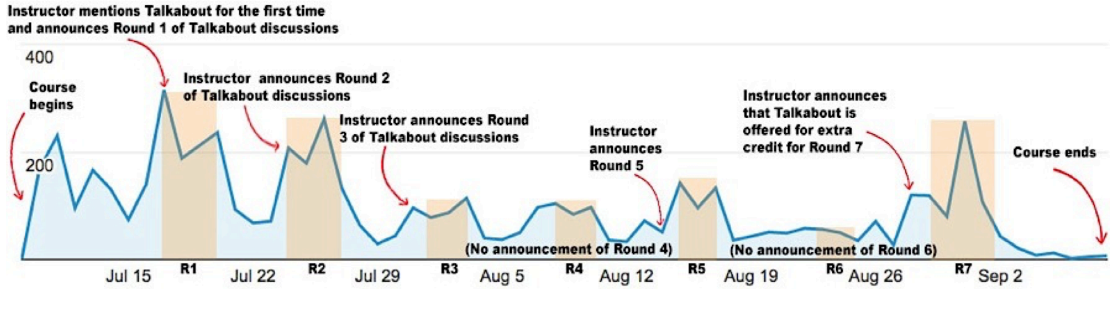


Figure 51: When instructors highlight peer-learning software, students use it. Talkabout pageviews of a women’s rights course. Instructor announcements are followed by the largest amount of Talkabout pageviews throughout the course. R1 represents Round 1 of Talkabout discussions, and so on, with orange rectangles framing the duration of each round. When instructor does not mention Round 4 and 6, pageviews are at their lowest.

The Talkabout course with the highest follow-through rate not only offered Talkabout for extra credit, but also offered technical support, including a course-specific Talkabout FAQ (Talkabout has an FAQ but it is not course specific). Looking at the forums, the role of the FAQ became apparent: many students posted questions about their technological difficulties and the community TAs and even other students would direct students to this FAQ – loaded with pictures and step by step instructions to help these students understand what Talkabout is and how it's related to them. Moreover, the course support team answered any questions that could not be answered by the FAQ, ensuring that anyone who was interested in using the peer learning platform got the chance to do so.

Going forward, online classes can also consider ways to accommodate students with differing constraints from around the world. For instance, Talkabout is not available to some students whose country (like Iran) blocks access to Google Hangouts. Other students may simply lack sufficient reliable Internet bandwidth. One course offered small-group discussions for credit that were held either online (with Talkabout) or in-person in order to combat this challenge. When the strongest incentives are impractical, courses can still improve social visibility to encourage participation.

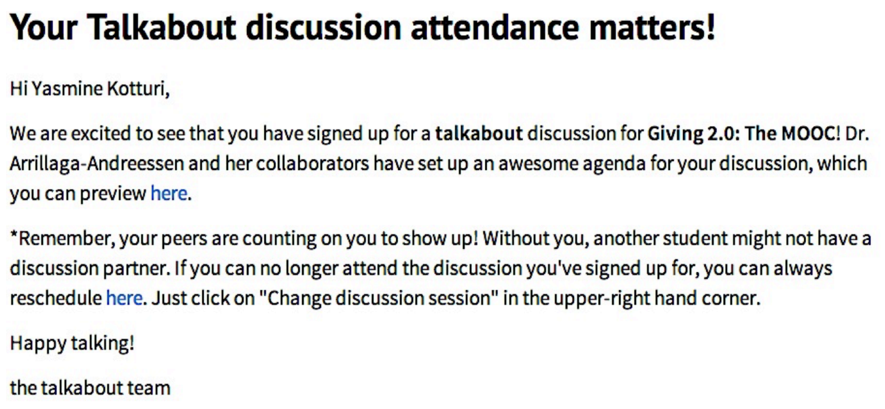
Effects of limited social translucence online

Online students are "hungry for social interaction" [247]. Especially in early MOOCs, discussion forums featured self-introductions from around the world, and students banded together for in-person meet-ups. Yet, when peer-learning opportunities are provided, students don't always participate in pro-social ways; they may neglect to review their peers' work, or fail to attend a discussion session that they signed up for. We asked 100 students who missed a Talkabout why they did so. 18 out of 31 responses said something else came up or they forgot. While many respondents apologized to us as the system designers, none mentioned how they may have let down their classmates who were counting on their participation. This observation suggests that social loafing may be endemic to large-scale social learning systems. If a student doesn’t feel responsible to a small set of colleagues and the instructor instead diffuses that responsibility across a massive set of peers, individuals will feel less compunction to follow through on social commitments.

To combat social loafing, we might reverse the diffusion of responsibility by transforming it onto a smaller human scale. Systems that highlight co-dependence may be more successful at encouraging pro-social behavior [250]. In a peer environment, students are dependent on each other to do their part for the system to work. Encouraging commitment and contribution can help students understand the importance of their participation, and create successful peer learning environments [245].

Norm-setting in online social interaction

Norms have an enormous impact on people’s behavior. In-person, teachers can act as strong role models and have institutional authority, leading to many opportunities to shape behavior and strengthen and set norms. Online, these opportunities diminish with limited social visibility, but other opportunities appear, such as shaping norms through system design. Platform designers, software and teachers can encourage peer empathy and mutually beneficial behavior by fostering pro-social norms.



Your Talkabout discussion attendance matters!

Hi Yasmine Kotturi,

We are excited to see that you have signed up for a **talkabout** discussion for **Giving 2.0: The MOOC**! Dr. Arrillaga-Andreessen and her collaborators have set up an awesome agenda for your discussion, which you can preview here.

*Remember, your peers are counting on you to show up! Without you, another student might not have a discussion partner. If you can no longer attend the discussion you've signed up for, you can always reschedule here. Just click on "Change discussion session" in the upper-right hand corner.

Happy talking!

the talkabout team

Figure 52: An email sent to students prior to their discussion, reminding them of the importance of their attendance, increases follow-through rate 41%.

Software can illuminate social norms online. For instance, PeerStudio reminds them of the reciprocal nature of the peer assessment process when students provide scores without written feedback (Figure 53).

As a different example, students that are late to a Talkabout discussion are told they won't be allowed to join the discussion, just as they'd not like to have a discussion interrupted by a late classmate. Instead, the system provides them an option to reschedule. Systems need not wait until things go wrong to set norms. From prior work, we know students are highly motivated when they feel that their contribution matters [238], [251]. As an experiment, we emailed students in two separate Talkabout courses before their discussion saying that their peers were counting on them to show up to the discussion (Figure 52). Without a reminder email, only 21% of students who signed up for a discussion slot actually showed up. With a reminder email, this follow-through rate increased to 62%.

How could software and students together highlight co-dependence?

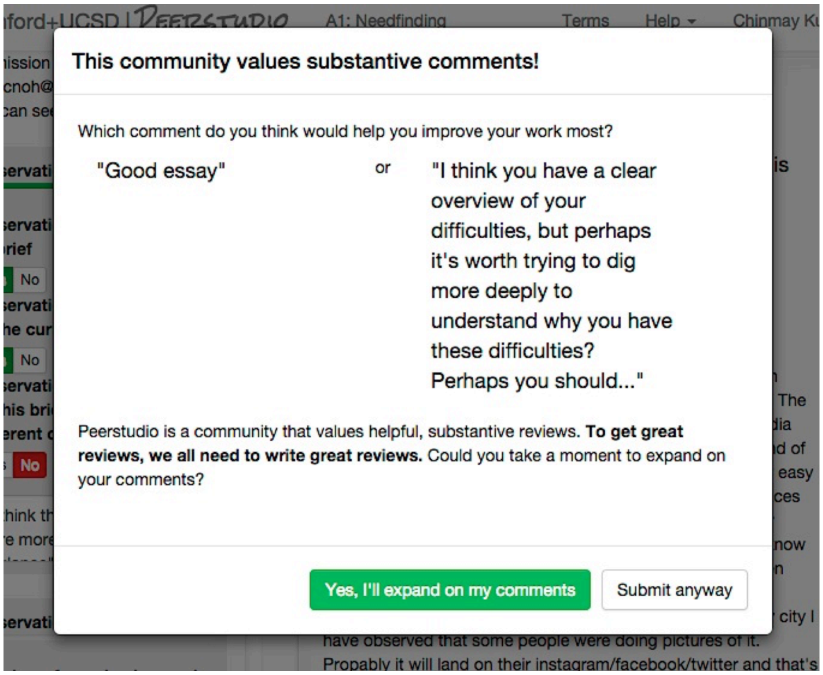
When few students are online, PeerStudio recruits reviewers by sending out emails to students. Initially, this email featured a generic request to review. As an experiment, we *humanized* the request by featuring the specific request a student had made. Immediately after making this change, review length increased from an average of 17 words to 24 words.

Humanized software is not the only influencer: forum posts from students sharing their peer learning experiences can help validate the system and encourage others to give it a try. For example, one student posted: "I can't say how much I love discussions…and that's why I have gone through 11-12 Talkabout sessions just to know, discuss and interact with people from all over the world." Although unpredictable [252], this word-of-mouth technique can be highly effective for increasing stickiness [253]. When students shared Talkabout experiences in the course discussion forums

(2000 posts out of 64,000 mentioned Talkabout, 3%), the sign-up rate was 6% (2037 students), and the follow-through rate was 63%. However, the same course offered a year later, did not see similar student behavior (260 posts out of 80,000 mentioned Talkabout, 0.3%). The sign-up rate was 5% (930 students) and follow-through rate was 55%. Although influenced by external factors, this suggests that social validation of the systems is important.

Leveraging students’ desire to connect globally

Increasing social translucence has one final benefit: it allows students to act on their desire for persistent connections with their global classmates. For example, incorporating networking opportunities in the discussion agenda allocates times for students to mingle: “Spend five minutes taking turns introducing yourselves and discussing your background.” However, we note that this is not a “one-size-fits-all” solution: certain course topics might inspire more socializing than others. For instance, in an international women’s rights course, 93% of students using Talkabout shared their contact information with each other (e.g. LinkedIn profiles, email addresses), but in a course



This community values substantive comments!

Which comment do you think would help you improve your work most?

"Good essay" or "I think you have a clear overview of your difficulties, but perhaps it's worth trying to dig more deeply to understand why you have these difficulties? Perhaps you should..."

Peerstudio is a community that values helpful, substantive reviews. **To get great reviews, we all need to write great reviews.** Could you take a moment to expand on your comments?

Figure 53: When PeerStudio detects a review without comments, it asks the reviewer if they would like to go back and improve their review by adding comments.

on effective learning, only 18% did.

Designing and hosting interactions from afar

Like a cook watching a stew come to a boil and adjusting the temperature as needed, an instructor guiding peer interactions in-person can modulate her behavior in response to student reactions. Observing how students do in-class exercises and assimilating non-verbal cues (*e.g.*, enthusiasm, boredom, confusion) helps teachers tailor their instruction, often even subconsciously [126].

By contrast, the indirection of teaching online causes multiple challenges for instructors. First, with rare exceptions [254], online teachers can't see much about student behavior interactively. Second, because of the large-scale and asynchronous nature of most online classes, teachers can't directly coach peer interactions. To extend – and possibly butcher – the cooking metaphor, teaching online shifts the instructor from the in-the-kitchen chef to the cookbook author. Their recipes need to be sufficiently standalone and clear that students around the globe can cook up a delicious peer interaction themselves. However, most instructors lack the tools to write recipes that can be handed off and reused without any interactive guidance on the instructor's part.

Writing recipes: scaffolding peer interactions from afar

Instructors using Talkabout early on often provided both too little student motivation and discussion scaffolding, and usage was minimal [234]. These unstructured discussion likely did not increase students' academic achievement or sense of community [248]. However, amongst these early instructors, we noticed that discussions when they did succeed, specifically targeted opportunities for self-referencing, highlighted viewpoint differences using boundary objects, and leveraged students as mediators [234].

Recipes are more successful if coupled with incentives

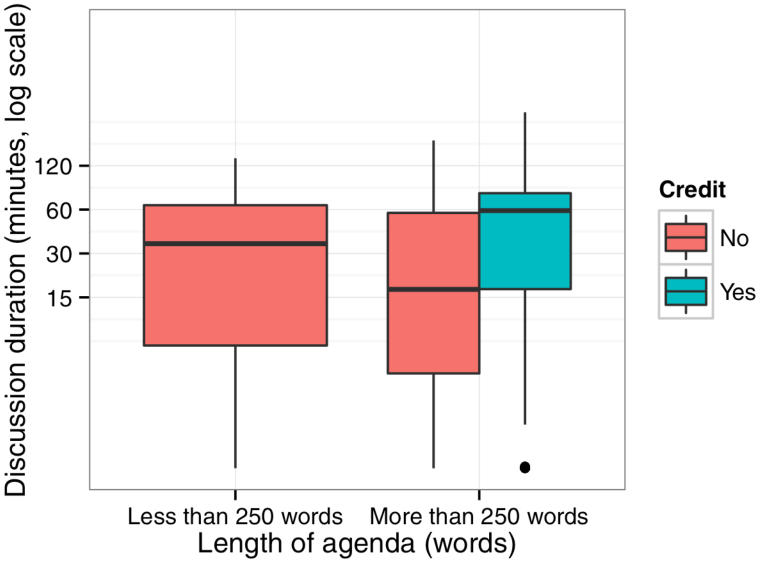


Figure 54: Longer discussion agendas incentivize students to discuss longer, but only when they are accompanied by course credit for participation.

Could other instructors successfully adopt these strategies? Before we built tools that might help instructors adopt these strategies, we wondered if discussion guidance alone predicted how long students discussed. To our surprise, we found that this is not the case: discussion guidance is only improves engagement (i.e. how long students discuss) if it is coupled with participation incentives, like course credit.

We split discussions into two categories: long and short discussion agendas, with 250 words as the threshold, and compared credit-granting and no credit discussions (Figure 54). All agendas asked students to discuss for 30 minutes. On average, students discussed for 31 minutes when given short agendas. Lengthier agendas had no effect without credit: the average length of discussion was 30 minutes. However, those long agendas that awarded credit successfully incentivized students to discuss longer: the average discussion with credit was 49 minutes.

Software can focus instructor attention where it is most valuable

A massive online classroom is unfamiliar territory to instructors trained to teach in the physical classroom. This unfamiliarity leads instructors to focus their energy on finely

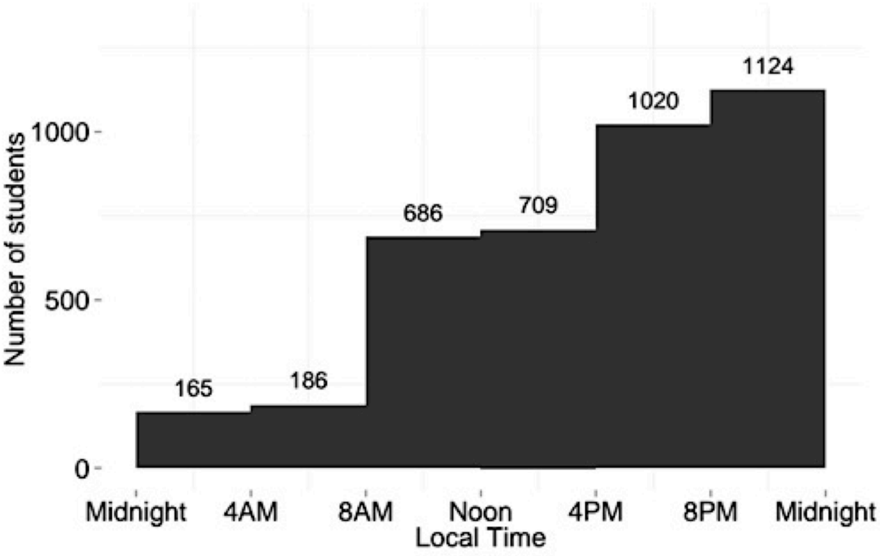


Figure 55: Most students discuss in the evening, but there are students that will discuss at all 24 hours (this graph shows data from nine classes and 3400 students.)

crafting some aspects of interaction, while potentially ignoring other more important aspects.

For example, time zones are a recurring thorn in the side of many types of global collaboration, and peer learning is no exception. Every Talkabout instructor was concerned about discussion session times and frequency, as this a major issue with in-person sections. Instructors often asked if particular times were good for students around the world. Some debated: would 9pm Eastern Time be better than 8pm Eastern Time, as more students would have finished dinner? Or would it be worse for students elsewhere? Other instructors were unsure of how many discussions timeslots to offer. One instructor offered a timeslot every hour for 24 hours because she wanted to ensure that there were enough scheduling options. However, an unforeseen consequence of this was that the participants were too spread out over the 24 discussions, and thus some students were left alone.

Our data, however, suggests that this effort would be better spent scaffolding discussions with effective agendas instead: scheduling is simply less important. Most students prefer evenings for discussions. Yet, different students prefer different times, with every time of day being preferred by someone (Figure 9). In summary, these data reflect how teachers' experiences in in-person classrooms do not translate easily online, and how tools can focus their attention on what does matter. We next consider how we could support instructors while they work on what does matter.

Teaching teachers by example

Even fantastic pedagogical innovation can be hamstrung when there is a mismatch between curricular materials and platform functionality. When curricula did not match to the needs of the setting, the learning platforms languished. We emphasize the importance of *teaching by example*: creating designs and introductory experiences that nudge instructors toward the right intuitions. While always true with educational innovation, the online education revolution is a particularly dramatic change of setting, and instructor scaffolding is particularly important.

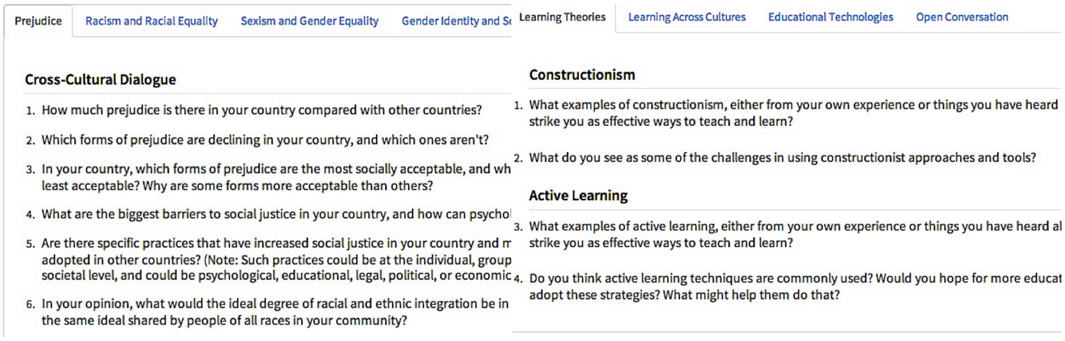
One of the most robust techniques we have found for guiding instructors is to provide successful examples of how other teachers have used the learning platform. In many domains, from design to writing research papers, a common and effective strategy for creating new work is to template off similar work that has a related goal [255]. During interviews with Talkabout instructors, a common situation recurred: the instructor was having a hard time conceptualizing the student experience. Therefore, as an experiment, we walked an instructor through Talkabout – in a Talkabout – and showed an excellent example agenda from another class. This helped onboard the new instructor to working with Talkabout: she was able to use the example as a framework that she could fill in with her own content (Figure 56).

We have also found example it useful to show instructors course announcements that described Talkabout using layman's terms with pictures of a Talkabout discussion. In addition to helping instructors think of why students should participate in Talkabout in their own class, these examples also spare them from having to describe the Talkabout system themselves. Most online students view course announcements, so a straightforward description can have large benefits.

One final way we have found examples to be valuable is to show instructors student behaviors that were otherwise invisible. For example, we showed an instructor a video clip of a Talkabout discussion along with a full discussion summary. In response, the instructor said, "The most interesting point was around the amount of time each student spoke. In this case, one student spoke for more than half of the Talkabout. This informs us to be more explicit with time allocations for questions and that we should emphasize that we want students to more evenly speak." By helping her visualize the interactions, she was able to restructure her discussion prompts in order to achieve her desired discussion goal; in this case, encouraging all students to have equally share their thoughts.

Conclusion

This chapter provided evidence for three challenges to global-scale adoption of peer learning, and offered three corresponding socio-technical remedies. We reflect on our experience from developing, designing and deploying our social learning platforms: Talkabout and PeerStudio, as well as our experience as teachers in physical and online classes. We looked at student practices, teacher practices and material design, and



Cross-Cultural Dialogue

1. How much prejudice is there in your country compared with other countries?
2. Which forms of prejudice are declining in your country, and which ones aren't?
3. In your country, which forms of prejudice are the most socially acceptable, and wh least acceptable? Why are some forms more acceptable than others?
4. What are the biggest barriers to social justice in your country, and how can psychol
5. Are there specific practices that have increased social justice in your country and m adopted in other countries? (Note: Such practices could be at the individual, group societal level, and could be psychological, educational, legal, political, or economic
6. In your opinion, what would the ideal degree of racial and ethnic integration be in the same ideal shared by people of all races in your community?

Constructionism

1. What examples of constructionism, either from your own experience or things you have heard strike you as effective ways to teach and learn?
2. What do you see as some of the challenges in using constructionist approaches and tools?

Active Learning

3. What examples of active learning, either from your own experience or things you have heard al strike you as effective ways to teach and learn?
4. Do you think active learning techniques are commonly used? Would you hope for more educat adopt these strategies? What might help them do that?

Example prompt **Instructor-created prompt**

Figure 56: Good agendas for Talkabout provided students a choice of questions for each topic, as well as a number of topics to choose from. Given examples of previous, successful discussions, instructors created prompts of their own that incorporated these best practices.

assessed the relationship between those and peer learning adoption. When peer systems and curricula are well integrated, the social context is illuminated, and teachers' and system designers' intuitions for scaffolding are guided by software, students do adopt these systems.

Acknowledgements

We are grateful to Scott Plous, Ed Hutchins and all instructors who used Talkabout and PeerStudio for their thought-provoking discussions and feedback; Coursera and Open EdX teams for their help integrating the platforms; the many students who interacted with their peers and shared their stories with us. This research was funded in part through NSF grants #1351131 and #1444865 and the Hasso Plattner Design Thinking Program. This research was conducted under the Stanford IRB protocols #30319 and #30324.

Chapter 8
Unforeseen side effects of global scale classrooms

The preceding chapters describe how software and pedagogy, when co-designed, achieve a desired learning goal through peer interactions. However, these interactions could sometimes also have unexpected side effects. This chapter discusses three examples of surprising side effects we have encountered. In hindsight, we were surprised by these side effects because we did not fully understand the nature of the online classroom. By reporting on them, we hope to help systems builders consider aspects we initially ignored, to encourage other researchers to report their failures, and to contribute to a fuller understanding of the online classroom.

Patriotic Grading

When we deployed our peer assessment system across many online classes, we observed that even though assessment was double blind (unless students explicitly wrote in their name in their submission), students assessed peer work from their own country higher than that from other countries. For example, in the HCI class, students in the first iteration of the class scored work from their own country a mean of 3.6% higher than work from other countries. Our data suggested this bias was not restricted to students residing in particular countries.

A patriotic grading bias penalizes students in countries with few peers, who are rated mostly by peers in other countries. Ideally, we would like students to rate work all peer work uniformly, so that students receives fair credit for their work. Understanding the reasons for a patriotic bias may help us design better assessment and create better scaffolding that might help students see work of distant peers more fairly. Furthermore, understanding this bias may also contribute to an understanding of global learning. For example, people in different parts of the world have differing intuitions of creativity and good design [256], and students may systematically prefer to solve problems using well-defined rules, or by analyzing each instance of a problem separately depending on where they live [199]. Could patriotic grading help us understand these norms better?

We studied patriotic biases by analyzing existing assessment data and through a series of controlled experiments. Our analysis of existing data suggest that patriotic biases are the result of students interpreting assignment instructions or submissions differently or of an implicit bias that causes students to rate work by "in-group" peers higher. We then explored if changing superficial details of students' submissions reduced this bias through controlled experiments. These experiments suggest that the bias is robust to superficial changes, and is perhaps caused by the actual content of student work.

Analysis of existing data: prevalence of bias and potential causes

Patriotism is present across classes and cohorts

We first observed a "patriotic bias" was in the first two iterations of Human-Computer Interaction class. Students in the first iteration of the class scored work from their own country a mean of 3.6% higher than work from other countries, and 3.1% in the second iteration. Curious to see whether this held for other courses and domains, we found a similar effect in *Listening to World Music*, by the University of Pennsylvania. In this class, students scored work from their country a mean of 1.8% higher, t(25095)=5.17, p<0.01.

While the magnitude of the bias varied across classes, these observations suggest that the bias itself was not limited to a particular class or domain; rather, this bias might be an effect of aspects of peer assessment that are common across classes. We considered three causes in particular: 1) students in different parts of the world understood assignment/assessment instructions differently 2) students were unable to understand their distant peers' work correctly because it used unfamiliar language or lacked local relevance or 3) this bias was due to social-distance, with students subconsciously rating 'in-group' peers higher.

Differing understanding of instructions cause bias, better rubrics may help reduce it

The bias in the second iteration of the HCI class was much smaller, even through the demographics of the two iterations were strikingly similar (Figure 57). We wondered whether redesigning rubrics and assignments caused this reduction.

In the third iteration of the class, we revised rubrics again, this time specifically looking for and eliminating colloquialisms, adding example work from outside the United States, and improving clarity in response to student questions. In this third iteration, students scored work from their own country a mean of only 1.5% higher than that from other countries; t(58987)= 6.7, p<0.01. While not a controlled experiment, this suggests that one potential reason for the bias was differently interpreted assignment descriptions, which can be reduced with better rubrics.

Just like modifying assignments was correlated with a lower bias, we wondered, would modifying student submissions have a similar effect? We tried two kinds of modifications: first, we tried to reduce an "in-group" bias by reducing overt signals

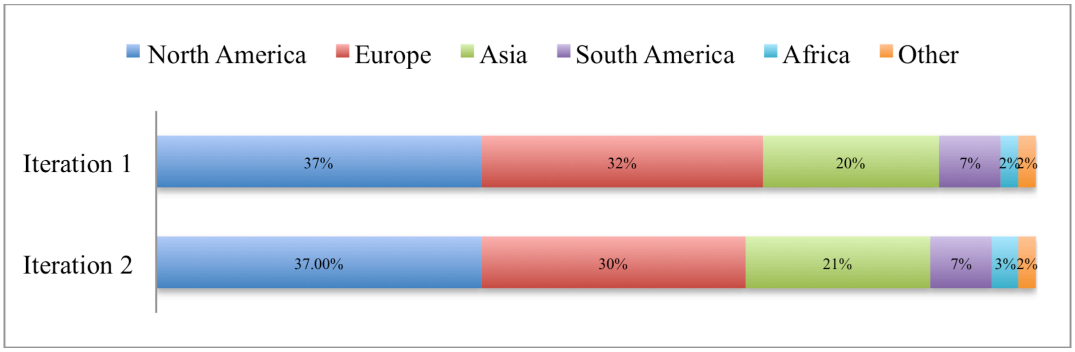


Figure 57: The first two iterations of the online HCI class attracted students in remarkably similar proportions from different geographical regions (more demographic data in Chapter 3.)

that students lived in a particular country. Second, we tried to make submissions understandable more uniformly around the world by reducing colloquialisms and language flaws.

Patriotism signaling study: Would removing overt signals of group belonging reduce bias?

In this experiment, we manually removed overt signals of origin by altering student submissions. We randomly selected 50 student submissions, and manually changed the names of every place mentioned in the submission to be a generic description. For example, for a student submission from the San Francisco Bay Area, "San Francisco" would be replaced by "nearby city", "Foothill Community College" was replaced by "local college". In addition, we replaced personal names by generic descriptors. For example "My friend, Joe" was replaced by "my friend", "I observed my sister Maria" became "I observed my sister". We focus on removing signals of strong group membership instead of making student work appear "in-group" because different raters assess the same work around the world, and our current experimentation platform doesn't allow showing a different version of work depending on the viewing student. If successful in reducing bias, this redaction could be automated based on Named Entity Recognition, which now works robustly even with noisy data [257].

We now describe an experiment that compares peer biases for these depersonalized submissions with the original submission.

Participants

This experiment was conducted in the third iteration of the Human-Computer Interaction class. Students could consent to sharing their peer assessment data when they submitted an assignment (roughly a week before peer assessment started); we report on the results of 1,558 students who consented.

Method

This experiment was conducted in the second and third assignments of the class. We randomly sampled from the student submissions on these assignments. We then inserted one submission from our randomly chosen subset into every student's grading set. Students either saw the original submission, or the modified version with names removed. This results in a between-subjects setup with two conditions (*Original* and *Depersonalized*). Students rated these submissions with the exact same interface and instructions as other peer submissions.

Manipulation check

Since we were removing information from student submissions, we wanted to see if this perceptibly reduced their quality. We found that staff (TAs) blind to condition did not rate the modified submissions differently than the original ones: t(28)=1.04, p>0.2.

Results: Depersonalization did not reduce bias

To analyze the results, we built a linear model predicting the raters' grade with two variables: the experimental condition (original or depersonalized), and a variable indicating if the rater and the original student whose work was shown to them were from the same country. With this model, if students' biases were based on signals such as names, we would see a significant interaction effect for the experimental condition.

However, we did not find a significant interaction effect: t(2344)=1.2, p~0.2; the bias was similar for both the original and depersonalized submissions. Suppressing overt signals did not reduce rating bias. The next experiment explores more subtle modifications.

Patriotism language fluidity study: Would increasing language fluidity reduce bias?

While altering submissions for the previous experiment, we noticed that submissions from non-English speaking countries had several language dis-fluencies. For example, a Spanish student who submitted work may have first written their submission in

Spanish and translated with Google Translate, resulting in non-idiomatic language (e.g. “They declare to be happy” instead of the more common “They reported to be happy”). This language could suggest low-quality work due to the Halo effect (see Chapter 4), or it could suggest an out-group student to native English speakers. Furthermore, just as assignment instructions may have been differently interpreted, student submissions might too. For example, a British student spoke about people using smartphones while walking on the “pavement”; American peers would likely be more familiar with “sidewalk.” We hypothesized that normalizing such language, in addition to cues relating to names of people and places, could reduce bias.

Participants

Participants in this experiment were the same as the previous experiment. We conducted this experiment on the fourth and fifth assignment in the class.

Method

The method was largely similar to the previous experiment. In this experiment, to prepare our modified submissions, we manually replaced colloquialisms with their generally accepted American counterparts and increased fluency by re-writing sentences that were grammatically incorrect or used non-idiomatic language (most commonly, this was due to machine translation translating proverbs and metaphors literally). In addition, we replaced names of people and places with generic forms, as described in the previous experiment.

Manipulation Check

In this experiment, we hope to improve the language used in student submissions. Therefore, if our manipulation is successful, the modified assignments should score higher. Performing a manipulation check similar to the previous experiment, we found that TAs, blind to condition, indeed rated our modified submissions higher: $t(29) = 1.99$, $p<0.05$.

Results: Language fluency and depersonalization did not reduce bias

Similar to the Patriotism Signaling study, we created a linear regression model to analyze results. Similar to the previous experiment, the interaction term would account for the benefit of the assignment modification, and a main effect of the experimental condition would account for any perceived increased quality of our modified submission.

As expected from the TA ratings, students rated the modified assignments higher: t(31943)=2.04, $p<0.05$. Unfortunately, we found no significant interaction effect for the experimental condition: foreign rater biases were statistically undistinguishable for original and modified submissions.

Overall, our experimental results suggest that reducing the assessment bias will take more than a superficial rewriting of student submissions. One could consider, for example, students working in pairs that span countries. Perhaps these interactions will allow the students to develop a more global understanding of class topics, and reduce the regional cues in submitted work.

Other biases: gender and income

Our discussion so far centers around rater biases based on students' countries of residence, but other rating differences also exist, such as when peers from rich and poor neighborhoods rate each other. In addition there are differences in how men and women rate their own work.

Patriotism SES study: Assessment Bias correlated with Average Neighborhood Income

For this Patriotism SES study, we look at how the economics of the neighborhoods students live in affects grading. We consider assessment data from the United States, which attracts the largest fraction of MOOC learners, and for which data about income distributions is publicly available. The United States also has amongst the world's largest levels of economic inequality [258]. Average income levels in a neighborhood, especially in the US, are correlated with education achievement, racial makeup, and even longevity [259]–[261]. Could these differences in students' neighborhoods also correlate with how students rate each others' work?

Method

We compared rater biases with differences in the average household incomes between the neighborhoods where students lived. We obtained zip-code level income data from the 2010 US Census, and combined it with students' approximate location based on IP addresses. While imprecise, IP location approximates demographic data well [262]. In all, we found locations for 684 students who submitted assignments in the HCI class and lived in the US. We discarded data for six students who lived in zip codes where the median income was more than two standard deviations away from the average student in our dataset. We then built a logistic regression model that predicted a student's rating of work given the difference in median incomes in the zip codes where the rater and submitter logged in.

Results: Income differences correlate with grading bias

We find that the absolute difference between the median incomes of the submitter's and rater's zip code predicts the grading bias: t(986)= -2.09, $p<0.05$. On average, every $10,000 in difference in median income reduced the peer grade by 0.46% of the total assignment grade (Figure 58). To put this in context, a student-rater pair from Sunnyvale, CA (median income $105,600) and Oakland, CA (median income $29,100) would rate each other 3% lower than a comparable pair between Sunnyvale, CA and Cupertino, CA (median income $122,400). While this income-differential rating bias is striking, it is still smaller than the international bias we encountered earlier. However, economic differences between countries are larger than those within each country (See Chapter 6 on Talkabout), so a similar causal mechanism may underlie both within-country and international biases.

Grading bias across genders

We consider two questions about differences in grading across genders: do men and women rate their own work differently? And do men and women rate peers' work differently?

Method

We asked students for their gender as part of the second iteration of the online HCI class; 2,594 students reported it (89.7% of all submitters): 1,071 female, 1,507 male, 16 other. Since a large fraction of students volunteered their gender, we only look at grading behavior amongst these 2,594 students. We built linear regression models that predicted the peer-awarded grade based on the students' self-assessed grade and their gender.

Here, we only report results for students who marked male or female because of the small fraction of students marking "other".

Results

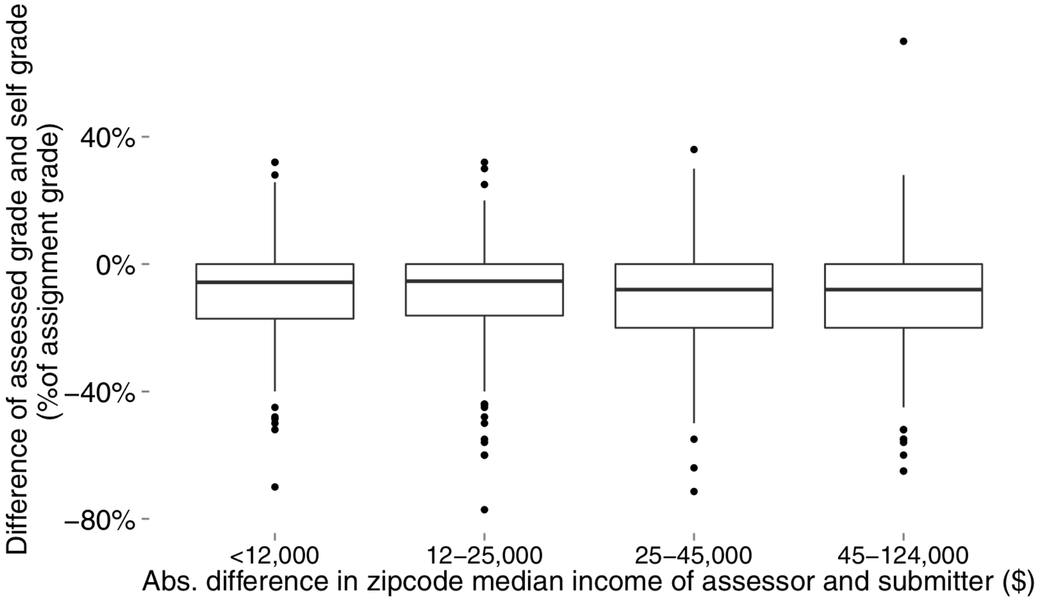


Figure 58: When double-blind, students within the US rate work higher when it is from neighborhoods with median incomes close to their own.

Compared to the peer-awarded grade, male students rate their own work higher through the self-assessment step than female students do: t(51383)=23.2, p<0.01. At the same peer-awarded grade, the average man rates their work 3.5% higher than the average woman does. Men also rate other students' work lower during assessment: t(46884)=3.9, p<0.01; on average men rate work 0.8% lower.

We also found that when double blind, both men and women rate work of submitters of both genders equivalently.

Comparison to an in-person class

Because the online HCI class is based on an on-campus class at Stanford, it is possible to roughly compare grading behavior in both classes. In the 160-person on-campus under-graduate class (http://cs147.stanford.edu), students do not rate each other, but they do self-assess, and are blind-rated by TAs. Compared to the TA grade, we found no statistical difference between how male and female students self-assess. This could be because the larger size of the online class provides the statistical power necessary to detect smaller differences (For an 80% chance to detect a 3.5% difference with our grade distributions, 1,500 observations are necessary). TAs in the on-campus class also trained students to assess work in person, which may have helped students rate their work more accurately.

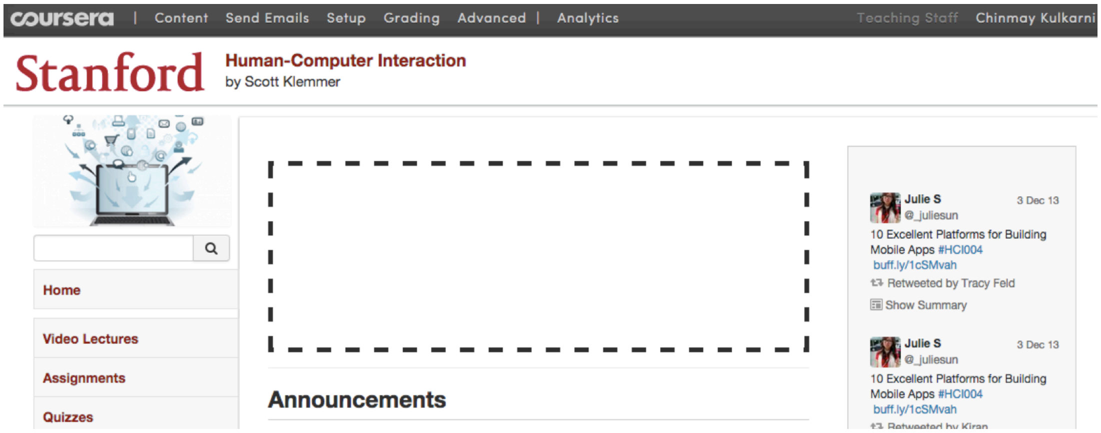
So far in this chapter, we have seen how students from different demographics (based on gender, location, income etc.) participate in a specific class activity—peer assessment. Next, we ask if we can increase student participation and persistence.

Attempts at improving persistence

Hundreds of thousands of students sign up for typical massive online classes, but a very small fraction of students complete. For example, only 7,100 students of the 155,000 enrolled in the online 6.002x MIT class (Circuits & Signals) completed it [125]. While some students see online classes as a way to engage their curiosity [35], many others intend to complete the class and put in sizeable effort over many weeks, but still fail to do so [40].

Could we help motivated students persist longer and succeed more frequently in classes? In dealing with this problem, we draw inspiration from research in online communities. Like online classes, online communities fail to keep newcomers who have put in sizable effort [245], [263]. To combat this attrition, many communities have introduced many commitment mechanisms for e.g. [251], [264]–[266]. Could we design tools for online classrooms that leverage mechanisms of social proof, and self-consistency biases that help spur people to action in other online communities? Could these tools provide some of the same commitment devices and social mechanisms present in physical classrooms? We designed and tested three such tools in 2012. Unfortunately, our tools failed to improve persistence, and in fact, our interventions significantly reduced it!

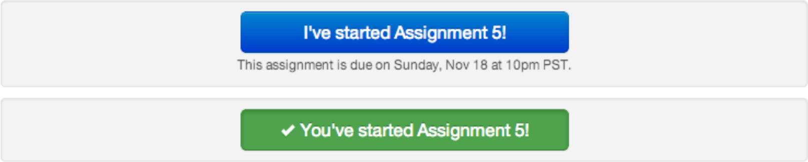
Self-consistency bias influences people to take actions which they believe to be consistent with their self-image [267]. Social proof refers to the effect of people heuristically assuming that others' actions are the correct behavior, especially in ambiguous situations [268].



Julie S @_juliesun 3 Dec 13
10 Excellent Platforms for Building Mobile Apps #HCI004 buff.ly/1cSMvah
Retweeted by Tracy Feld
Show Summary

Julie S @_juliesun 3 Dec 13
10 Excellent Platforms for Building Mobile Apps #HCI004 buff.ly/1cSMvah
Retweeted by Kiran

Figure 59: Experimental tools showed at the top of the student's course homepage (replacing the dashed box above.)



This assignment is due on Sunday, Nov 18 at 10pm PST.

Figure 60: The Recording tool. (Top) Default state, inviting student to report they've started work. (Bottom) The recorded state is shown to students who click the button.

Informed by this research, we designed systems that we hoped would encourage students to persist in class. Our *Recording* tool allowed students to click a button ("I've started on this assignment"), and displayed this information to them prominently to encourage self-consistency. The *Proof by Friends* tool was similar in terms of making students' actions salient, except that students would also connect it to Facebook so a student's Facebook friends would see this progress via email and when they logged on to the course web site. The *Proof by Crowd* tool replaced the need to connect students to Facebook, and would instead broadcast their action more broadly to all students in the class that were using it. Through the Proof by Friends and Proof by Crowds tools we hoped that seeing peers start work on an assignment would create social proof that working on an assignment was expected behavior.

Unfortunately, our intervention had unintended consequences. In a randomized controlled study, students used one of the three tools, or were assigned to the control condition (no tool). As expected, students who clicked the button signifying that they'd started an assignment were self-consistent and did usually finish the assignment. However, displaying this button also caused other students in the *Recording* condition to neither click the button nor submit the assignment. Overall, fewer students in the *Recording* condition submitted assignments that students in the no-button control condition. One possible explanation for this result is that students in the *Recording* condition had their intrinsic motivation to complete assignments crowded out by the extrinsic motivation of a commitment/policing mechanism.

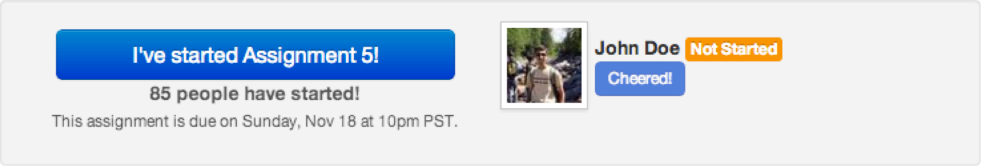
Showing students their friends' behavior also did not improve completion rates above the control condition either. We hypothesize that this was because very few students had more than one Facebook friend also enrolled in class. Seeing the actions of a single Facebook friend (in the Proof by Friends condition) may not have created enough social proof. It is also possible that strangers were simply not persuasive enough to create social proof in the *Proof by Crowd* condition, similar to other research [269].

While these results represent a failed system, our hope is that they inspire other researchers to pay more careful attention to how intrinsic and extrinsic motivations interact. Perhaps they will also inspire other researchers to carefully consider issues of critical mass while deploying systems that rely on social effects. Overall, this experience suggests that tools are more likely to work when they encourage existing motivations, and provide students a clear benefit, even with a small cohort of users.

Using self-consistency and social proof to improve persistence

Self-consistency bias influences people to take actions which they believe to be consistent with their self-image [267]. Thus, even a small action such as writing down that you will complete a goal will later make that person more likely to follow through [88]. The Foot-in-the-door persuasion technique [250] leverages this bias, by first asking a person for a small favor, which makes them more likely to perform a subsequent, larger demand later. For example, members of online communities who share their goals publicly are more likely to meet their goals [265]. Perhaps closest to this work, requiring students to sign a learning contract leads to better performance and attitudes [270].

Social proof refers to the effect of people heuristically assuming that others' actions



85 people have started!

This assignment is due on Sunday, Nov 18 at 10pm PST.

Figure 61: The Proof by Friends condition. The Proof by Crowd condition was identical except it did not show the list of friends on the right.

are the correct behavior, especially in ambiguous situations [268]. Users have a higher probability of taking an action when they see others have done so [271].

To examine whether these effects could help students persist in class, we performed a randomized, controlled experiment that compared the completion rates of students using our tools (each of which embody one of these effects) with a control group of students who did not have access to them.

Participants

This experiment ran during the Fall 2012 offering of the Coursera Human-Computer Interaction class (www.hci-class.org). In the class, students had between one to two weeks to complete each of 5 open-ended assignments. Each assignment took about 5 to 10 hours to complete. The experiment was advertised on the Fall 2012 class website as an opt-in, experimental social feature.

At the start of the class, 2384 students signed up for the experiment. Students who signed up consented to data collection about their grades and assignment-completion. They were not compensated and did not receive extra course credit for participation. Over the course of the experiment, 28 students withdrew. Of the rest, 305 students had at least one Facebook friend who also participated in the experiment. We only include these 305 students in our analysis.

Experimental setup

This experiment had four between-subjects conditions. In the control condition, students experienced the class normally, except for a notice on the class homepage that informed them that they were on a wait-list for social features in class. In the *Recording* condition, students saw a button ("I've started on this assignment") on the class homepage and assignment page (Figure 60). The system would record when students clicked it and display this information on the two pages, along with a notice that that this information was private and not shared with classmates. In the *Proof by Crowd* condition, students saw the “I’ve started” button, and the total number of students in class who had started the assignment. In the *Proof by Friends* condition, students saw the information in the *Proof by Crowds* condition as well as a list of their friends along with whether they had started the assignment (Figure 61). All tools were shown prominently on the course homepage, when students were logged in (Figure 59).

In all conditions, students received two emails for every assignment: a reminder a few days before an assignment was due, and a reminder to peer-assess other students right after an assignment due date passed. To make the social proof more salient, in the *Proof by Friends* condition, the assignment reminder emails included which friends had started the assignment; the peer assessment reminder emails included a link their friends' assignments. Correspondingly, in the *Proof by Friends* condition, the assignment reminder email included the number of students nearby (within 100 miles) who had started the assignment.

Experimental assignment: To combat the chance that students in the control condition might feel excluded from additional class features (and thus demotivated), experiments in this chapter used a wait-list control, also commonly used in medical studies [272]. We informed all participants about the features we were introducing, and put everyone on a waitlist, ostensibly so we could test features with a small set of students before rolling them out. Students in the control condition remained on the waitlist, while we assigned the rest to other conditions. To ensure that groups of friends were treated similarly (e.g. all friends could see each other on the web site), we performed simple network bucketing, assigning whole groups of friends randomly to the chosen condition.

Analysis

Because students were exposed to this system across multiple observations (assignments), we built a logistic mixed-effects model that predicted whether a student would submit an assignment. The experimental condition was a fixed-effects term, and the model had a random intercept for each student. We also included a random intercept term for each assignment, as assignments had varying difficulty, and might have different completion rates. Because we didn't expect the experimental methods to have differing efficacy for different assignments, we did not include an interaction term. We report results from this linear time-independent model for simplicity of interpretation; a time-dependent survival analysis model yields the same results.

Results

Recording hurts persistence

Students in the Recording condition were 46% less likely to submit an assignment than those in the waitlist condition: $t(296) = -3.91$, $\beta=-0.86$, $p< 0.05$.

Using a model built only with students in the *Recording* condition, our data also suggest clicking the “I've started” button indicates a strong intent to complete the assignment: those that click the button are more than twice as likely to complete the assignment than those that don't, $\beta = 5.26$, $t(74)=8.3$, $p<0.05$.

Recording might hurt overall persistence, but does it make students who do submit assignments work harder? We found the opposite. Using the word length of submissions as a proxy for effort, we found that students wrote 200 fewer words (approximately 22.5% of an average submission) in the Recording condition than the control condition.

Social proof mechanisms did not improve persistence

Students in the *Proof by Crowds* and *Proof by Friends* condition were not significantly more likely to submit assignments: $t(296)= -0.34$, $p > 0.5$, and $t(296) = 0.45$, $p >0.5$ respectively.

Limitations of results

This experimental design assumes symmetry: in the *Proof by Friends* and *Proof by Crowd* you can both say you've started an assignment and see who else has. While this yields a simple mental model for participants, it does not allow us to separate the effect of self-consistency and social proof. Due to limitations of the class platform, we could not create a system that automatically detected when a student had started working on an assignment. Future work that incorporates such an automatic update might find techniques based on social proof to be valuable.

The experimental social platform had other features in the spirit of increasing social proof. In particular, students could "cheer on" their friends by clicking a button. Using Facebook messages, students could also give their friends feedback about their work, as well as encourage friends who did not complete the assignment to continue in the class. Fewer than ten students used these features, so we do not think they affect our primary results.

Discussion

While we find evidence of self-consistency, we also find that a mechanism that requires students to perform an action to record their progress reduces the overall number of students completing an assignment.

We see two possible explanations for this finding. First, students could have viewed the "I've started" button as a policing mechanism from staff. Since they voluntarily signed up for the class, such policing would be inconsistent with their class goals, and make the goal of completion extrinsically rewarded, crowding out their higher, intrinsic motivation [273]. Second, it may be that the Recording condition promised no obvious benefit – information about starting an assignment was not shared with others in the class. Once students had made a decision to not click the button, their self-consistency bias may have worked in the opposite direction, making them less likely to complete work.

We were also surprised with the lack of improvement in persistence due to social proof. In a post-experimental survey, a participant in a non-social condition wanted to "be able to see how [his friends] were performing", while another without Facebook friends lamented that "there's no one to show off your progress to". Our data suggests that students do find such information valuable. One student wrote, "only one of my friends was taking the lesson (class)", but she "was able to see what he has done and that was good." Given such feedback, it is surprising that seeing friends' progress is not motivating.

Again, we see two potential explanations. First, students may have seen too few friends to create a perceived norm: fewer than 50 students in our class had more than a single friend enrolled. Another explanation is that social proof is a two-edged sword. One participant had "a couple of friends on the course[;] they quickly stopped doing assignments (and frankly so did I)." We find this explanation less likely, because we saw no beneficial statistical trends even when students had friends that completed work.

Overall, these results suggest that researchers must be mindful of the crowding effects of their differing motivational schemes, and motivational tools should emphasize pre-existing motivations.

We now turn to attempts to improve student achievement directly. Our data, as well as the experience of other researchers, shows that students who aren't performing well tend not to persist [35]. Therefore, in the following study, we hope to increase both student achievement and performance.

Do examples of good peer work improve students' performance?

Instructors often stress the inspirational benefits of peer work. Guidelines for many project-based classes suggest that instructors show examples of great student work, to set norms and inspire students. On the other hand, novices are less successful at transferring ideas from examples that are distant than examples that are more proximal. This suggests that students who perform poorly on an assignment might learn better by seeing examples of work that is just a little better than their own, rather than the inspiring examples of their most successful classmates. In a large, online classroom, is the norm-setting benefit of a great example dominant, or the ease of transferring from more similar work? And could we get the best of both worlds, by showing students both kinds of examples?

To determine which examples, the ones with the highest quality, or the most similar in quality, help the most, we conducted a randomized controlled experiment in the HCI class on Coursera (Chapter 3 offers more details about the class composition and structure).

This experiment targeted two research questions:
RQ1: Do students who see examples of either high-quality work or work that is accessible but better than their own score higher in future class assignments?
RQ2: Do students who see examples of either high-quality work or work that is only slightly better than their own persist for longer in class?

The results of this experiment suggest that examples we showed did not improve student tenure in class. Furthermore, the examples that we showed students did not improve their performance; on the contrary students who saw excellent examples did significantly worse in their future work than students who did not see any examples. The following sections detail the experimental setup and analysis that led to this intriguing result.

Participants

We conducted this experiment in the second iteration of the HCI online class, in January 2013. In all, 1573 students consented to participation and submitted at least two assignments. Because of our goal is to measure the effect of examples on future performance, we exclude students who submitted fewer than two assignments.

Experimental setup

This experiment used a between-subjects setup, where students either saw two examples of excellent peer work, which scored in the 95th percentile or higher on the assignment (the excellent-examples condition), two submissions of work that were graded no more than 5% better than their own, but less than 10% better than own work (the "approachable examples" condition), or two examples, one approachable and one excellent (the "mixed examples" condition). In the control condition, students saw a message ("We are unable to find you examples of peer work to learn from at this time") instead of the examples ("no examples" condition). In all experimental conditions, students were told to compare examples to their own work, and find ways to improve it. We hoped such explicit comparisons would reveal structural attributes of better work, and lead to greater abstraction [274].

In designing this experimental setup, we wanted to balance the accuracy of experimental manipulation with applicability to a large online class, where student work is typically peer assessed. Therefore, we only selected examples where all peers agreed on the grade to within 5% of the assignment grade. This results in most participant seeing examples that have the desired quality without sacrificing the applicability of our results to the real world.

Procedure and measures

After the end of the peer-review period for each assignment, the class website showed students appropriate examples along with their assignment grade. Students could return to this webpage any time until the end of the class. Students saw the same examples every time.

Instructors consider assignment scores as valid measures of work quality, and students are motivated to perform well on assignments through certificates of achievement. Therefore, to measure the quality of work, we used the student's grade in an assignment. Assignments were peer assessed. Because these assignments require significant student effort, we used the number of assignments a student submitted over the length of the course as a dependent measure of persistence.

Analysis

We use assignment grades as a measure for work quality. In analyzing our data, we built a linear regression model that predicted a student's grade in an assignment given their grade in the previous assignment and their experimental assignment (adjusted R^2 = 0.104).

We measure persistence through the number of assignments a student submits. We built both a simple linear regression model, and a more complex survival analysis model to analyze our data. Both models have similar outcomes, so we describe results from the regression model, which are easier to interpret. This model predicts the number of assignments a student submits based on their experimental assignment. We found this model (and the survival analysis) to have very poor model fit (adjusted R^2 = -0.0008).

Results: Excellent examples reduce future work quality

Compared to student in the no-examples condition, we found that students in the excellent-examples condition scored lower, though the effect was only marginal, $t(1101) = -1.81$, $p = 0.07$. On average, students in the excellent-examples condition scored 2.7% lower. Students in approachable- and mixed- example conditions did not perform significantly differently from those in the no-example control condition ($p > 0.4$).

Results: Examples don't inspire persistence

In our experimental setup we did not find any effect of experimental assignment and student persistence.

Discussion

Taken together, our results suggest that examples of peer work didn't improve students' work; rather, seeing excellent examples hurt the quality of future work. We see two potential explanations. First, unlike experts (such as teachers in the classroom), novices may not be able to understand which aspects of their peers' work they could adopt into their own [274], [275]. Therefore, without additional scaffolding, students may try and emulate their peers, but may instead adopt superficial features. Informal feedback from students supports this notion. Students frequently observed for example that the examples we chose were "more elaborate", but didn't comment on what specific attributes they'd like to adopt.

Second, examples may reduce students' motivation, by making the gap between their own work, and that of well-performing peers more salient. This lack of motivation may then reflect in their future work.

More generally, this experiment shows both the promise and the current limitations of completely automated educational systems. While such systems may find examples based on pre-defined criteria from the large corpus of student examples, they may not provide the scaffolding that makes these examples effective.

Implications for design

This chapter discusses three unforeseen side effects of massive-scale peer interactions. Initially, my view of these results was as simple failures: the unfortunate price of doing evidence-based design that resulted both in frustrated collaborators and a diminished learning experience for students.

However, this simple view disregards three important aspects. First, these unforeseen side effects probably operate at all scales, but are only detectable at the massive scale of an online class. For example, our work on motivation might be applicable to small classrooms as well, where failed attempts by to improve persistence through commitment mechanisms might go unnoticed. Second, large scale enables not only to detect these subtle effects, but allows for evidence-based methods to combat them. For example, in our work on peer assessment, we found that simple revising and clarifying rubrics can reduce geographic rating bias. Finally, these side effects may help us understand the current limits of large-scale instruction. For example, we found that simply showing students examples of excellent work was insufficient. Perhaps, future work may reveal that instructors' explanations of what makes work excellent is necessary as well.

Going forward, we offer three suggestions to researchers. First, consider the normative motivations of the environment. At present, many online students are intrinsically motivated. As credentials count for more, this could change, making extrinsic motivators more valuable. Second, consider how tools may be useful at different levels of student adoption, and when possible, design systems so that they are valuable at a wide range of adoption. For example, the *Social Proof* system described here lacks effectiveness when students have few friends. In contrast, PeerStudio achieves fast reviewing even when few students are online by recruiting students over email as a backup. Finally, consider how students using your tools simultaneously from around the world might be beneficial. For example, in contrast to the rater-bias discussed in this chapter, Talkabout capitalized on geographic diversity through its targeted discussion guides.

Chapter 9
Contributions and future directions

This thesis exploits the networked properties of online classrooms and extends the benefits of peer learning to massive scale. It does so through systems that decide whom learners interact with amongst thousands of diverse peers, and with interfaces that scaffold these interactions, while simultaneously dealing with the large size of the class, and the asynchronous and remote access in an online class.

Contributions to structuring peer interactions for education

This thesis shows how creating opportunities for massive-scale peer interactions can improve learning and engagement.

- This dissertation research provides the first analysis of massive-scale peer review. It describes a system where students are first trained on assessment using rubrics and a staff-curated set of submissions, and then each student independently rates peer work based on a rubric and accompanying examples. We find that the average difference between peer and staff grades using this peer assessment system was 6.7%, and a median of five peer ratings was within 10% of the staff 65% of the time. This dissertation also shows that providing students feedback on their grading accuracy improves this accuracy even further. By examining common errors students make, we then describe how instructors can create "fortune cookies" that enable peers to provide actionable, qualitative feedback. Finally, we show that students rate peers from their country significantly higher than peers from other countries, even though assessment is double-blind.

- This dissertation then introduces a method to combine peer assessment with machine classifiers to focus students' assessment effort where their judgment is most valuable. This method uses a machine classifier's prediction confidence to determine the number of assessors assigned to each submission. Through testing in a live class, we show that combining peer and machine assessment in this way results in 80-90% of the accuracy of peer-only assessment with only 54% of student effort.
- This thesis introduces PeerStudio, a system for fast revision-oriented feedback. PeerStudio uses the potential temporal overlap between student schedules in a large online class and the resulting temporal overlap in accessing the system. We describe how such overlapping access can be used to recruit peers quickly. In an online class with 572 students submitting work, the median time to recruitment was 7 minutes. This thesis also introduces the "back-off" recruitment method that recruits reviewers by email when few are available online, while minimizing the number of students it emails to ask for help. Finally, we demonstrate how rapid feedback from PeerStudio improves student performance (by 4.4% on an essay question in our evaluation).
- PeerStudio also incorporates interactive machine generated hints that help students write better reviews; analysis of data suggests that these hints result in more actionable feedback.
- This thesis introduces Talkabout, a system that leverages global participation in online classes to create small-group discussions with diverse participants and contrasting perspectives. In a series of controlled experiments, we show that when students discuss in groups where participants are drawn from many countries, their grades improve by as much as 6% (in *Irrational Behavior*). We also introduce techniques for developing discussion guides that encourage students to share diverse experiences and perspectives.

- Through a retrospective analysis of our data, we show how visible instructor involvement in a peer learning intervention such as emails to students that encourage them to participate, or personally participating in such activities is correlated with large voluntary adoption of such opportunities. Similarly, closely aligning the design of peer learning interventions with pedagogical goals is also correlated with large adoption.
- This thesis also describes that when peer-learning systems run counter to existing student motivation, such as when they crowd out intrinsic motivations through policing or commitment mechanisms, can have negative effects. Similarly, simply mimicking the form of a classroom technique (e.g. showing examples of excellent work) does not achieve the learning goal of the activity. Instead, the actual mechanism of learning must be captured as well (e.g. instructors explaining the examples).

Impact and recent developments

Informed by the work presented in this dissertation, researchers have advanced the state of large-scale online education software and pedagogy. Below, we discuss a subset of such work. We only include work where we have personally interacted with the authors, or work that directly cites papers this thesis is based on. Being more familiar with such work enables us to reflect upon the challenges of research in this field, and suggest promising paths for future work.

Improving peer assessment

The design of the peer assessment system described in Chapter 3 was driven by the desire to create a general widely applicable assessment method that did not use any knowledge about the domain, assignment, or students. Therefore, this system relied on 1) rubrics, which are widely applicable and 2) an un-weighted median to aggregate peer ratings, which doesn’t incorporate knowledge about domain or students.

Other researchers have since explored other peer assessment schemes incorporate information about student work, or incentivize students for more accurate assessment, and systems that that do not use rubrics.

Improving assessment accuracy with more information

Using information about raters and assignments may improve assessment accuracy beyond the system from Chapter 3. We deployed one such system, developed by Chris Piech, in the third iteration of the online HCI class [14]. This system improves assessment accuracy significantly and requires no input from the instructor, and its use resulted in a large improvement in grade agreement with instructors. However, improved accuracy comes at the cost of complicating the mental model of learners, and many pass/fail find the un-weighted aggregation has sufficient accuracy. As a result, systems to improve assessment accuracy have not been widely deployed. As classes offer more nuanced credentials, systems that improve accuracy may be adopted more widely.

One opportunity to improve assessment is to use more information about peer raters. Raters rate work predictably higher or lower than the consensus score [14]. Some students also more reliably rate submissions closer to their consensus score than others [14]. Therefore, Piech et al propose a Bayesian model that starts assuming every student rates similarly (this is the implicit assumption in our system). Then, it updates each rater's *bias* and *variance* by observing the rating behavior of each student across multiple submissions for the same assignment. These two properties correspond to each of their observations respectively. Using this method allows them to approximate staff assessment much more accurately.

Another opportunity in improving peer assessment accuracy is to leverage information about the work itself. For example, the Codewebs system for programming assignments relies on homework submissions being highly structured, and propagates peer/teacher feedback through the system by finding similar programs [62]. Similarly, CaptainTeach combines unit tests for programming with peer assessment [276].

Incentivizing accurate assessment

Our work suggests that students can assess more accurately with feedback on their grading accuracy (Chapter 3). Lu et al. show that part of this benefit may come from students knowing their work is being monitored: when students know that their own peer assessment efforts are being assessed by peers, they more reliably distinguish between good and bad work. When students assess their peers' assessment effort, this effect is amplified [277].

However, incentive mechanisms may also inadvertently nudge students to assess according to what they believe the consensus to be, rather than their own independent judgment This can be especially problematic as students acquire expertise, and more successfully assess deeper aspects of work that novices miss [78]. A promising solution is to use the Bayesian Truth Serum technique [278], which asks raters both to provide their own independent assessment of work, and predict what the consensus assessment would be. Beyond improving assessment accuracy, Bayesian Truth Serums could also help identify experts; systems could also help experts so identified prevent the expert blind spot [279].

Different models of peer assessment

There are instances where scores and ranking are not required, and feedback is all that is necessary, e.g. on early drafts. However, soliciting numeric ratings in addition to general feedback induces peer reviewers to explain their comments in significantly greater detail, and make more positive comments [280]. This suggests that demanding a summative assessment focuses reviewers to engage critically with peer work. Having so engaged reviewers, how might we best focus reviewer attention?

The PeerStudio system asks raters to compare two pieces of work (one student submitted, one expert curated) along a number of rubric dimensions, and aggregates these comparisons to produce a numeric grade. This comparative method is faster than assessing work in isolation, and requires little training (recall that PeerStudio, unlike our system in Chapter 3, had no training step). Furthermore, comparisons are more discriminative among items when raters are deeply familiar with each item [281]. However, the speed of these methods also encourages raters to use cognitive shortcuts and heuristics [171], which may lead raters astray if they skip seeking critical information (see Chapter 4; Halo effect).

Other researchers have also since employed comparative ratings in peer assessment. In ordinal rating, raters perform a summative comparison between two peer submissions, e.g. [282], or they rank a small number of student submissions [283], typically along one dimension. These comparisons are then aggregated into an approximate total ordering of goodness of submissions. An open question for future work is how ordinal ratings across multiple dimensions can be aggregated into a single rating.

Applying peer assessment to creative crowd work

While the contributions in assessment discussed above are in the educational context, creative tasks on crowdsourcing markets also benefit when workers assess their own work or receive feedback [105] and when they assess others workers [284]. Greenberg et al find that similar to a class context, combining rubrics with examples enables crowd workers to acquire the micro-expertise necessary for writing high-quality critiques [15].

Discussions in massive online classes

Talkabout enables informal video discussions that are loosely structured through discussion guides, and occur synchronously; these design decisions are driven by Talkabout's goal to leverage the diversity in an online class (Chapter 7). Other researchers have addressed other pedagogical goals with other design decisions.

MoocChat uses synchronous structured text discussions for teaching logical thinking and problem solving. `

Talkabout demonstrates how video discussions with a loosely enforced structure improve learning through diversity in experience and points of view. In contrast, text chat systems such as MoocChat demonstrate how text discussions with a strict conversational structure, and enforcing equal contribution from participants improve learning, potentially through more critical thinking. Text chat also opens up opportunities for semi-synchronous conversations. For such conversations to improve learning, it will be necessary for both the speaker and listener to understand each other's contribution [56]. Pre-recorded video contributions successfully allow remote participants to contribute to a conversation, but conveying the discussion experience back to the participant remains an open question [285].

Future Directions

At their core, the systems in this dissertation leverage the global community of peers to improve learning in existing large online classes. Future work could build classes or even educational systems that employ peer interactions as their primary pedagogical asset. In particular, I see three opportunities for future work.

Creating new educational opportunities for a diverse world

Some critics argue that universities act as sieves [286], being selective in who they admit and bestowing upon them knowledge, skills, and social capital (leading to economic and social success). Could online classes offer a more emancipating alternative?

Selectivity is an asset in physical classrooms [287], but perhaps the collective expertise and experience of students online could be a greater asset online? Perhaps online classrooms could function not like sieves, but instead like cities, attracting diverse people, and creating value through opportunities for students to interact with peers they might not otherwise meet.

To better leverage diversity as a pedagogical asset, MOOCs could become more diverse as well. MOOCs may well be the most diverse classrooms so far, but much can be improved for equitable access. For example, students in Africa are underrepresented in MOOCs, and have a smaller completion rate compared to the rest of the world [35]. Even within the United States, MOOCs haven't been able to attract poorer students, even though they may benefit more from MOOCs [288]. Our data suggest that the average MOOC student lives in a neighborhood with a median income of $72,000 per year, much higher than the median for the United States ($51,000). Encouraging students with more diverse experiences to participate may make tools such as Talkabout even more valuable, and will be crucial if we want to design online classes like cities, where the collective experience of the participants is a core pedagogical asset.

Combining global and local interactions

One the one hand, Talkabout demonstrates how global classmates can improve learning at massive scales by exposing students to contrasting perspectives and making implicit assumptions salient. On the other hand, the jigsaw classroom demonstrates that in collocated environments familiarity can lead to greater trust and empathy among diverse students and improve learning [29]. Could systems combine the benefits of collocated familiarity and global perspectives be combined? For example, systems could enrich the familiar jigsaw discussion by "teleporting" students to distant locations?

Creating learners prepared for future learning

In a physical classroom, enabling lifelong learning has remained challenging because it is expensive and disruptive to bring students back to a physical classroom to teach new topics; it is much cheaper to deliver education “in bulk” [289]. In contrast, approximately 60% of students in online classes enroll because they want to keep up with new developments in the field, or want to apply knowledge from the class in their job [40]. Could the experience of these students suggest it may finally be possible for online classes to enable lifelong learning effectively?

To fully grasp this opportunity for lifelong learning, teaching methods and software could prepare students for future learning [290]. For instance, peer interactions could be designed to maximize transfer of knowledge to new situations, for example requiring students to discuss and invent a solution before it is revealed in class [7]. Similarly, peer interactions could train students in activities that are critical to future learning, such as negotiation, planning team goals, persistence, and reacting to critical feedback [290].

Appendix 1
Transcript of a Talkabout

This is a transcript of a Talkabout session in the Organizational Analysis class. The transcript skips the first few moments of the session, before participants consented to recording. The discussion agenda asks students to discuss one of two potential topics for 30 minutes. This group, like many others, chose to discuss both (~60 minutes).

Aditi[10]: Do we want to go with organizational culture or with organizational learning? Personally, I vote, if we have time, let's do both. So, if you can go ahead, um

Mark: Basically, they are very, they are connected.

Everyone: Mhmm

Mark: I think we could talk about the both.

Aditi: Okay, so would anyone of you like to start with an example of organizational culture that you thought was very prominent and you could see the difference between two?

Tiffany: I will. I am actually experiencing a transition in my workplace right now. This current staff feels like the culture is in jeopardy. We are losing our manager who hired each and every one of us um and over the course of the last four years, we've just really developed a consistent attitude among the staff. She hired people with good sense of humor and people who enjoy technology and love learning and those are the people that stuck around. I'm very weary about how the new manager will take over

Tiffany continues to talk about her situation for 5 more minutes

Mark: So you have changed the culture in your organization?

Tiffany: Yeah

Mark: That's great. That example is really great, I think.

Jose: But now her question is, but, uh, let me ask you another question just a small one. Do you have anxiety about the new person coming in?

[10] Names throughout this chapter have been changed to protect student privacy.

Tiffany: Right

Jose: But you aren't sure that he -- It's very much possible that it will go in the same way as the things are going on now.

Tiffany: Yeah

Jose: How are you deciding? [How are you making a decision about what your future boss will be like?]

Tiffany: It's too early to decide what is it going to be and I think that is the part that gives me a little bit of anxiety. I've been told many times just be patient and here are some things about this manager that will blend really nicely with the group and everything will work out just fine.

***Aditi** agrees*

Olga: I have a question: is it possible for you to change the organizational culture at the library?

Tiffany: Not at our particular branch because we are so small, um each person really has an impact on the culture. Whenever it is hostile or negative that is eventually resolved. But at the larger branch that we are a part of its been a huge problem. Things are imminently slower, and super, super resistant against change.

Olga: Thank you

Tiffany: Does anyone else have any examples in their workplace?

Tiffany and **Mark** go back and forth for around 5 minutes

Aditi shares an example from her experiences living in India

Mark, Tiffany, Jose, and **Aditi** continue to share their opinions

Bill chimes in after being quite this whole time.

Aditi invites **Yuree** to share her thoughts since **Yuree** hasn't said anything yet

Yuree: Hi.

Aditi: Hi Yuree

Yuree: Uh yeah. Actually I heard thoughts of discussion. I'd liked to share my organization learning in my experiences which is *[I've worked in]* American companies, *[and]* I've worked in China *[in]* two departments, R&D and marketing. My work focuses on the learning and development and how to collaborate between the two departments, I think that is one of the questions of organizational learning.

Yuree continues for three minutes about her situation

Aditi: During this process, did you have any difficulty? Did they have any problem with the learning and the merging?

Yuree: Uh yes. People have pressure to work over time *[poor connection, hard to understand]*

Aditi: Excellent. Thank you Yuree. Does anyone have any other questions for Yuree?

No response

Aditi: Thank you Yuree. Amy, would you like to add anything?

Amy: Actually what I was going to say what that I was going to agree with Bill. I've worked at Hewlett Packard for over 20 years and I've been involved in several acquisitions. I can assure you there is no such thing as a merger, there are only acquisitions. They actually buy and sell the culture of their company. Now the other thing --

Bill: That's what --

Amy: I'm sorry, go ahead --

Aditi: Bill, were you saying something?

Amy: Me? Yes? Can you hear me?

Aditi and Tiffany: Yes

Amy: Could you hear me before?

Aditi and Tiffany: Yes

Amy: Okay, but initially, when you are talking about people not wanting to embrace change because it is coming from the top down, you're quite right. The only way that change gets implemented in an organization is when the management team goes to the ground level, to the people at the bottom, and actually asks them what's going on, and they implement those changes.

Amy continues for two minutes

Olga shares her opinion for five minutes

Tiffany: If any has anything to share, I can start a forum with today's date

Mark: That would be nice

Tiffany: Okay

Aditi: Um actually, why don't you guys carry on with the discussion? I have another course that starts at 11 o'clock, right now, so I will take your leave. And I want to thank all of you for all of your valuable input and I always say, the more I talk in these Google Hangouts, it motivates more to do much better in the course. So again, thank you very much for all of your input.

Everyone says thank you

Aditi: Buhbye.

Bill: Actually its midnight here in Singapore so I think the rest of you in the world where the hour is more sensible so you are probably more alert than I am. Yuree, you are in China?

Yuree: Yes.

Bill starts speaking Mandarin

Yuree laughs and says something in Mandarin

Bill: Yeah I should probably check out, as we say, Yuree will now *[mandarin]*.

Mark: Woah, impressive! *[laughs]*

Yuree: Yes, this is a very valuable discussion, from being in the HR department. Yeah, that's all...it is late for me so I would like to discuss next time. Bye.

Mark: Have a nice evening, nice night.

Tiffany: In that case, au revoir.

Everyone says goodbye and leaves the Hangout

Appendix 2
Test materials used as measures for Talkabout

The Social Psychology class had a 50-question final. Scores on the exam were used as a measure of academic performance in Experiment 3. The exam was designed to “cover the lectures, assigned videos, and assigned readings”, “intended to be roughly as difficult as an average college-level psychology exam in the United States”. Below, we reproduce every tenth question to give a sense of topics covered.

Question 10

As discussed in the assigned reading, research by Brad Bushman and his colleagues (2009) found that people who were _______ in narcissism and _______ in self-esteem tended to behave more aggressively than did other people.

(a) low; high (b) low; low (c) high; low (d) high; high

Question 20

A meta-analysis by Rod Bond and Peter Smith (1996) found that the best predictor of conformity levels in Asch-style research was:

(a) Whether the participants were female or male (b) Whether the majority group was made up of ingroup or out-group members (c) The year when the study was conducted (d) Whether the culture was individualist or collectivist

Question 30

Zhang Wei becomes scared while watching a horror film about a murderer who hides outside people's houses. After the film, he may be more likely to interpret a sound outside his house as threatening because of:

(a) Self-monitoring (b) The misinformation effect (c) The foot-in-the-door phenomenon (d) Priming

Question 40

Research on post decisional dissonance suggests that on average, students will feel more confident of their answers on this exam before submitting the exam than after submitting it.

(a) True (b) False

Question 50

The out-group homogeneity effect occurs even when people have extensive contact with members of an out-group.

(a) True (b) False

We reproduce every fifth question from the first two weeks' quizzes from the Organizational Analysis class below. These quizzes were used as measures in Experiment 1 and Experiment 3. Note that these quizzes were independently created by the instructor in a previous run of the class (before Talkabout was designed), and were used unchanged in the experimental class.

Week 1 Quiz

Question 5

Which of the following are reasons why organizational theories are important (select all that apply)?

(a) They afford perspectives beyond your own individual experience (b) They allow you to better understand and interpret complex phenomena (c) They provide generalizable knowledge that can be useful in a variety of familiar and unfamiliar contexts (d) They can help you be a better manager (e) They explain everything that goes on in every organization in a way that makes things clear and simple

Question 10

Identify the corresponding class of organizational theory "The organization is thought to have multiple actors with potentially conflicting goals. These actors often form emergent and organic coalitions."

(a) Rational (b) Natural (c) Open

Question 15

Identify the corresponding class of organizational theory “Organizations are viewed less as making decisions and more as responding and adapting to their environment.”
(a) Rational (b) Natural (c) Open

Week 2 Quiz

Question 5

Which of the following are necessary in order to make fully or ideally rational decisions (select all that apply)?
(a) Knowledge of all your possible actions or choices. (b) Knowledge of the consequences associated with each possible action or choice. (c) Knowledge of your preferences. In other words, you need a way of ranking possible consequences in terms of their desirability. (d) More time, information, and attention than most people possess in most situations.

Question 10

The logic of appropriateness and the logic of consequences are equally concerned with the expected consequences of a particular action.
(a) True (b) False

Question 15

The second reform effort in the Chicago Public Schools was characterized by an emphasis on accountability, centralization of power, and Republican leadership.

(a) True (b) False

Appendix 3
Irrational Behavior Agenda

1. Are you irrational?

Are your parents? Friends? Enemies? Frenemies? What cases can you think of where the people around you exhibit some of the irrational tendencies that Dan describes in his lectures?

2. Subtle Influences.

What subtle influences in the consumer environment might have an effect on your purchases? What could you do to counteract these influences, or push your behavior in the desired direction?

3. Decision Illusions.

What "decision illusions" do you see in the real world? Do any current events come to mind where decision makers have been influenced by their environments?

4. Swaying Preferences.

What kinds of preferences do you think might be more stable than others? When can our decisions be swayed, and in which cases do we have a firm hold on our preferences?

5. Irrational Work.

How do the findings in behavioral economics relate to your area of study or work? Do you see irrationality in your workplace?

6. Cultural Differences.

How might cultural differences come into play with (ir)rational behavior?

7. Are defaults unavoidable?

When there must be a default option, how can we design defaults so that we encourage behavior optimally?

Do defaults save lives? Yes? No? Maybe so?

8. Humans and Economics.

What is the value of adding a human component to the field of economics? What are the possible advantages and disadvantages?

9. Choice Architecture.

What examples of “choice architecture” do you see in your own lives? Which choice sets or environments would you design differently?

10. Paternalism and Policy.

How much paternalism is too much? How should policymakers strike the right balance between encouraging proactive behavior in their constituents and safeguarding free will?

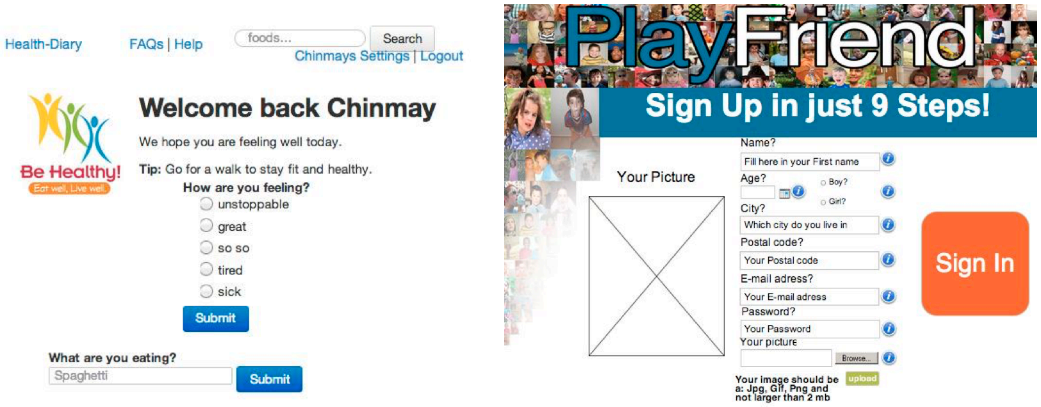
Appendix 5: Rating differences between staff and peers

Agreement between peer grades and staff grades without aggregation

Comparing the peer grades (not their medians) with staff grades demonstrates the value of aggregating peer grades (Figure 63). 26.3% of grades were within 5% of staff grades, and 46.7% within 10%. (Recall that the median agreement was 42.% and 65.5%, respectively)

Grading differences

Where peer graded higher: Figure 62(a) shows an application a student created as "an interactive website which helps people tracking their eating behavior and overall-feeling, to find and be able to avoid certain foods which causes discomfort or health related problems." Peers rated the prototype highly for being "interactive". Staff, rated it low, because "while fully functional, the design does not seem appropriate to the goal. The diary aspect seems to be the main aspect of the app, yet it's hidden behind a search bar."



Health-Diary FAQs | Help foods... Search
Chinmays Settings | Logout

Be Healthy!
Eat well, Live well.

Welcome back Chinmay

We hope you are feeling well today.

Tip: Go for a walk to stay fit and healthy.

How are you feeling?

○ unstoppable
○ great
○ so so
○ tired
○ sick

Submit

What are you eating?
Spaghetti
Submit

PlayFriend

Sign Up in just 9 Steps!

Your Picture

Name?
Fill here in your First name

Age?
○ Boy?
○ Girl?

City?
Which city do you live in

Postal code?
Your Postal code

E-mail adress?
Your E-mail adress

Password?
Your Password

Your picture
Browse...

Your image should be a: Jpg, Gif, Png and not larger than 2 mb
upload

Sign In

(a) Submission where peers grade higher than staff

(b) Submission with staff grade higher than peers

Figure 62: Student submissions with large differences between staff and peer grades.

Where peers graded lower Figure 62(b) shows an application a student created as an "exciting platform, bored children can engage (physically) with other children in their neighborhood." Staff praised it as "fully interactive, page flow is complete", while some peers rated it "unpolished", and asked the student to "Try to make UI less coloured."

Sample Rubric

Table V shows a rubric for the "Ready for testing" assignment. All other rubrics are available as online supplementary materials.

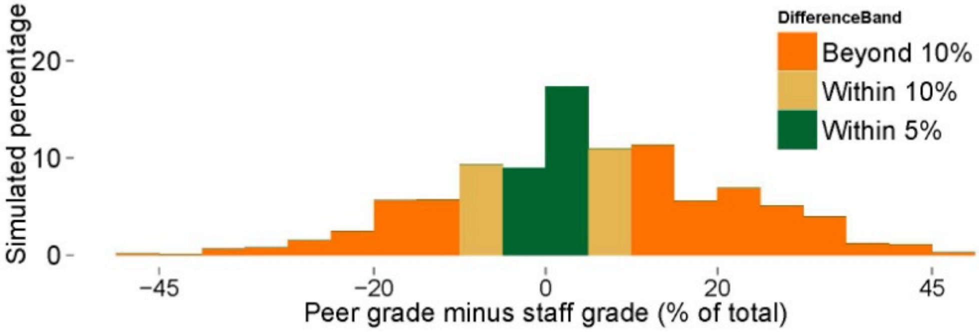


Figure 63: Agreement of un-aggregated peer grades and staff grades. Agreement is much lower than between median peer grades and staff grades.

References

[1] D. L. Schwartz, R. Lindgren, and S. Lewis, "Constructivism in an age of non-constructivist assessments.," in *Constructivist instruction: Success or failure?*, 2009, pp. 34–61.

[2] A. Pendleton-Jullian, *Four (+1) Studios*. CreateSpace Independent Publishing, 2010.

[3] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, "Active learning increases student performance in science, engineering, and mathematics," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8410–8415, 2014.

[4] P. L. Peterson and S. R. Swing, "Students' cognitions as mediators of the effectiveness of small-group learning.," *Journal of Educational Psychology*, vol. 77, no. 3, p. 299, 1985.

[5] R. E. Slavin, "Cooperative learning," *Learning and Cognition in Education Elsevier Academic Press, Boston*, pp. 160–166, 2011.

[6] C. Araragi, "The effect of the jigsaw learning method on children's academic performance and learning attitude.," *Japanese Journal of Educational Psychology*, 1983.

[7] D. Schon, *The Reflective Practitioner: How Professionals Think In Action*. 1983.

[8] H. G. Andrade, "The Effects of Instructional Rubrics on Learning to Write," *Current Issues in Education*, vol. 4, no. 4, 2001.

[9] H. G. Andrade, "Teaching with rubrics: The good, the bad, and the ugly," *College Teaching*, vol. 53, no. 1, pp. 27–31, 2005.

[10] T. Lewin, "One Course, 150,000 Students," 2012.

[11] D. A. Schön, *The design studio: An exploration of its traditions and potentials*. RIBA Publications for RIBA Building Industry Trust London, 1985.

[12] K. A. Ericsson and P. Ward, "Capturing the Naturally Occurring Superior Performance of Experts in the Laboratory: Toward a Science of Expert and Exceptional Performance," *Current Directions in Psychological Science*, vol. 16, no. 6, pp. 346–350, Dec. 2007.

[13] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, and others, "Crystal structure of a monomeric retroviral protease solved by protein folding game players," *Nature structural & molecular biology*, vol. 18, no. 10, pp. 1175–1177, 2011.

[14] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, “Tuned Models of Peer Assessment in MOOCs,” in *Proceedings of The 6th International Conference on Educational Data Mining*, 2013.

[15] M. D. Greenberg, M. W. Easterday, and E. M. Gerber, “Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers.”

[16] A. Parameswaran, S. Boyd, H. Garcia-Molina, A. Gupta, N. Polyzotis, and J. Widom, “Optimal Crowd-Powered Rating and Filtering Algorithms,” in *40th International Conf. on Very Large Data Bases (VLDB)*, 2014.

[17] M. Joglekar, H. Garcia-Molina, and A. Parameswaran, “Evaluating the Crowd with Confidence,” in *SIGKIDD*, 2013.

[18] K. A. Ericsson, R. T. Krampe, and C. Tesch-Römer, “The role of deliberate practice in the acquisition of expert performance.,” *Psychological review*, vol. 100, no. 3, 1993.

[19] C. Kulkarni, M. Bernstein, and S. Klemmer, “Rapid peer feedback in MOOCs emphasizes iteration and improves performance,” in *Learning at Scale (to appear)*, 2015.

[20] M. Stevens and M. Kirst, *Remaking College: The Changing Ecology of Higher Education*. Stanford University Press, 2015.

[21] P. Gurin, E. L. Dey, S. Hurtado, and G. Gurin, “Diversity and higher education: Theory and impact on educational outcomes.,” *Harvard Educational Review*, vol. 72, no. 3, 2002.

[22] R. Fry, “College enrollment hits all-time high, fueled by community college surge,” *Pew Research Center Publications*, 2009.

[23] C. Goldin and L. F. Katz, “Transitions: Career and family life cycles of the educational elite,” *The American Economic Review*, pp. 363–369, 2008.

[24] S. Okita, J. Bailenson, and D. Schwartz, “The Mere Belief of Social Interaction Improves Learning,” in *Cognitive Science Conference*, 2007.

[25] D. Tinapple, L. Olson, and J. Sadauskas, “CritViz: Web-Based Software Supporting Peer Critique in Large Creative Classrooms,” *Bulletin of the IEEE Technical Committee on Learning Technology*, vol. 15, no. 1, p. 29, 2013.

[26] D. Chinn, “Peer assessment in the algorithms course,” *ACM SIGCSE Bulletin*, vol. 37, no. 3, pp. 69–73, 2005.

[27] S. Dow, A. Kulkarni, B. Bunge, T. Nguyen, S. Klemmer, and B. Hartmann, “Shepherding the crowd: managing and providing feedback to crowd workers,” in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, 2011, pp. 1669–1674.

[28] D. Boud, *Enhancing learning through self assessment*. Routledge, 1995.

[29] E. Aronson and D. Bridgeman, “Jigsaw groups and the desegregated classroom: In pursuit of common goals,” in *Readings About The Social Animal*, Worth Publishers, 2004, p. 532.

[30] P. Pintrich and A. Zusho, “Student motivation and self-regulated learning in the college classroom,” *The scholarship of teaching and learning in higher education: An evidence-based perspective*, pp. 731–810, 2007.

[31] K. R. Wentzel, L. Filisetti, and L. Looney, “Adolescent prosocial behavior: The role of self-processes and contextual cues,” *Child Development*, vol. 78, no. 3, pp. 895–910, 2007.

[32] L. R. Antil, J. R. Jenkins, S. K. Wayne, and P. F. Vadasy, “Cooperative learning: Prevalence, conceptualizations, and the relation between research and practice,” *American educational research journal*, vol. 35, no. 3, pp. 419–454, 1998.

[33] L. S. Fuchs, D. Fuchs, J. Bentz, N. B. Phillips, and C. L. Hamlett, “The nature of student interactions during peer tutoring with and without prior training and experience,” *American Educational Research Journal*, vol. 31, no. 1, pp. 75–103, 1994.

[34] L. S. Fuchs, D. Fuchs, C. L. Hamlett, N. B. Phillips, K. Karns, and S. Dutka, “Enhancing students’ helping behavior during peer-mediated instruction with conceptual mathematical explanations,” *The Elementary School Journal*, pp. 223–249, 1997.

[35] R. F. Kizilcec, C. Piech, and E. Schneider, “Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses,” *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 170–179, 2013.

[36] L. B. Breslow, D. E. Pritchard, J. DeBoer, G. S. Stump, A. D. Ho, and D. T. Seaton, “Studying learning in the worldwide classroom: Research into edX’s first MOOC,” *Research & Practice in Assessment*, vol. 8, pp. 13–25, 2013.

[37] C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer, “Peer and self assessment in massive online classes,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 20, no. 6, p. 33, 2013.

[38] R. E. Slavin, *Student team learning: A practical guide to cooperative learning*. ERIC, 1991.

[39] S. Kiesler, R. E. Kraut, P. Resnick, and A. Kittur, “Regulating behavior in online communities,” *Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA*, 2012.

[40] R. F. Kizilcec and E. Schneider, “Motivation as a lens to understand online learners: Toward data-driven design with the OLEI scale,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 22, no. 2, p. 6, 2015.

[41] R. F. Kizilcec, “Collaborative learning in geographically distributed and in-person groups,” in *AIED 2013 Workshops Proceedings Volume*, 2013, p. 67.

[42] J. Hollan and S. Stornetta, “Beyond being there,” in *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’92*, 1992, pp. 119–125.

[43] F. M. Harper Yan Chen, J. Konstan, and S. X. Li, “Social comparisons and contributions to online communities: A field experiment on movielens,” *The American economic review*, pp. 1358–1398, 2010.

[44] C. Lampe and P. Resnick, “Slash (dot) and burn: distributed moderation in a large online conversation space,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 543–550.

[45] R. B. Cialdini, C. A. Kallgren, and R. R. Reno, “A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior,” *Advances in experimental social psychology*, vol. 24, no. 20, pp. 1–243, 1991.

[46] Y. Ren, R. Kraut, S. Kiesler, and P. Resnick, “Encouraging commitment in online communities,” *Building successful online communities: Evidence-based social design*, pp. 77–125, 2012.

[47] M. A. Hogg and J. C. Turner, “Interpersonal attraction, social identification and psychological group formation,” *European Journal of Social Psychology*, vol. 15, no. 1, pp. 51–66, 1985.

[48] D. Abrams, K. Ando, and S. Hinkle, “Psychological attachment to the group: Cross-cultural differences in organizational identification and subjective norms as predictors of workers’ turnover intentions,” *Personality and Social psychology bulletin*, vol. 24, no. 10, pp. 1027–1039, 1998.

[49] S. Hurtado, J. Milem, A. Clayton-Pedersen, and W. Allen, “Enacting Diverse Learning Environments: Improving the Climate for Racial/Ethnic Diversity in Higher Education. ASHE-ERIC Higher Education Report, Vol. 26, No. 8.,” Nov. 1998.

[50] T. F. Pettigrew, “Intergroup contact theory.,” *Annual review of psychology*, vol. 49, pp. 65–85, Jan. 1998.

[51] M. Hewstone, “Intergroup contact: Panacea for prejudice?,” *Psychologist*, 2003. [Online]. Available: http://www.psy.ox.ac.uk/publications/28661. [Accessed: 19-May-2014].

[52] R. T. Johnson and D. W. Johnson, “Cooperative learning in the science classroom,” *Science and children*, vol. 24, pp. 31–32, 1986.

[53] J. Piaget, *Piaget’s theory*. Springer, 1976.

[54] C. S. Symons and B. T. Johnson, “The self-reference effect in memory: A meta-analysis.,” *Psychological Bulletin*, vol. 121, no. 3, pp. 371–394, 1997.

[55] M. C. Wittrock, “Generative learning processes of the brain,” *Educational Psychologist*, vol. 27, no. 4, pp. 531–541, 1992.

[56] L. Devin-Sheehan, R. S. Feldman, and V. L. Allen, “Research on children tutoring children: A critical review,” *Review of educational Research*, pp. 355–385, 1976.

[57] S. Michaels, C. O’Connor, and L. B. Resnick, “Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life,” *Studies in philosophy and education*, vol. 27, no. 4, pp. 283–297, 2008.

[58] R. Kumar, C. P. Rosé, Y.-C. Wang, M. Joshi, and A. Robinson, “Tutorial dialogue as adaptive collaborative learning support,” *Frontiers in artificial intelligence and applications*, vol. 158, p. 383, 2007.

[59] M. A. Hearst, “Can Natural Language Processing Become Natural Language Coaching? (Keynote),” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2015.

[60] J. Kim, P. J. Guo, C. J. Cai, S.-W. D. Li, K. Z. Gajos, and R. C. Miller, “Data-driven interaction techniques for improving navigation of educational videos,” in *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 2014, pp. 563–572.

[61] E. L. Glassman, J. Scott, R. Singh, P. J. Guo, and R. C. Miller, “OverCode: Visualizing variation in student solutions to programming problems at scale,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 22, no. 2, p. 7, 2015.

[62] A. Nguyen, C. Piech, J. Huang, and L. Guibas, “Codewebs: scalable homework search for massive open online programming courses,” in *Proceedings of the 23rd international conference on World wide web*, 2014, pp. 491–502.

[63] M. Brooks, S. Basu, C. Jacobs, and L. Vanderwende, “Divide and Correct: Using Clusters to Grade Short Answers at Scale,” in *Learning at Scale*, 2014.

[64] T. Lewin, *Education Site Expands Slate of Universities and Courses*. The New York Times, 2012.

[65] J. Widom, *From 100 Students to 100,000*. ACM SIGMOD Blog (\url{http://wp.sigmod.org/?p=165}), 2012.

[66] B. Buxton, *Sketching user experiences: getting the design right and the right design*. Morgan Kaufmann, 2007.

[67] M. A. Hearst, “The debate on automated essay grading,” *Intelligent Systems and their Applications, IEEE*, vol. 15, no. 5, pp. 22–37, 2000.

[68] B. Lawson, *How designers think: the design process demystified*. Architectual Press, 2006.

[69] T. Winograd, “What can we teach about human-computer interaction?(plenary address),” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1990, pp. 443–448.

[70] S. Greenberg, “Embedding a design studio course in a conventional computer science program,” in *Creativity and HCI: From Experience to Design in Education*, Springer, 2009, pp. 23–41.

[71] J. E. Tomayko, “Teaching software development in a studio environment,” *ACM SIGCSE Bulletin*, vol. 23, no. 1, pp. 300–303, 1991.

[72] A. Drexler, R. Chafee, and others, *The Architecture of the Ecole des Beaux-Arts*. distributed by MIT Press, 1977.

[73] Y. J. Reimer and S. A. Douglas, “Teaching HCI design with the studio approach,” *Computer Science Education*, vol. 13, no. 3, pp. 191–205, 2003.

[74] B. Uluoglu, “Design knowledge communicated in studio critiques,” *Design Studies*, vol. 21, no. 1, pp. 33–58, 2000.

[75] K. Cennamo, S. A. Douglas, M. Vernon, C. Brandt, B. Scott, Y. Reimer, and M. McGrath, “Promoting creativity in the computer science design studio,” in *Proceedings of the 42nd ACM technical symposium on Computer science education*, 2011, pp. 649–654.

[76] D. P. Dannels and K. N. Martin, “Critiquing Critiques A Genre Analysis of Feedback Across Novice to Expert Design Studios,” *Journal of Business and Technical Communication*, vol. 22, no. 2, pp. 135–159, 2008.

[77] A. Snodgrass and R. Coyne, *Interpretation in architecture: Design as a way of thinking*. Routledge, 2006.

[78] T. M. Amabile, “Social psychology of creativity: A consensual assessment technique.,” *Journal of personality and social psychology*, vol. 43, no. 2, pp. 997–1013, 1982.

[79] E. B. Feldman, *Practical art criticism*. Prentice Hall New York, 1994.

[80] M. Tohidi, W. Buxton, R. Baecker, and A. Sellen, “Getting the right design and the design right,” in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006, pp. 1243–1252.

[81] D. Fallman, “Design-oriented human-computer interaction,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2003, pp. 225–232.

[82] J. Forlizzi and K. Battarbee, “Understanding experience in interactive systems,” in *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, 2004, pp. 261–268.

[83] L. Alben, “Defining the criteria for effective interaction design,” *interactions*, vol. 3, no. 3, pp. 11–15, 1996.

[84] R. E. Bennett, M. Steffen, M. K. Singley, M. Morley, and D. Jacquemin, “Evaluating an Automatically Scorable, Open-Ended Response Type for Measuring Mathematical Reasoning in Computer-Adaptive Tests.,” *Journal of Educational Measurement*, vol. 34, no. 2, pp. 162–176, 1997.

[85] R. E. Bennett, “Validity and automated scoring: It’s not only the scoring,” *Educational Measurement: Issues and Practice*, vol. 17, no. 4, 1998.

[86] S. Hsi and A. M. Agogino, “Scaffolding knowledge integration through designing multimedia case studies of engineering design,” in *Frontiers in Education Conference, 1995. Proceedings., 1995*, 1995, vol. 2, pp. 4d1–1.

[87] C. A. Stanley and M. E. Porter, *Engaging Large Classes: Strategies and Techniques for College Faculty.* ERIC, 2002.

[88] B. J. Zimmerman and D. H. Schunk, “Reflections on theories of self-regulated learning and academic achievement,” *Self-regulated learning and academic achievement: Theoretical perspectives*, vol. 2, pp. 289–307, 2001.

[89] P. R. Pintrich, “Understanding self-regulated learning,” *New directions for teaching and learning*, vol. 1995, no. 63, pp. 3–12, 1995.

[90] K. Topping, “Peer assessment between students in colleges and universities,” *Review of Educational Research*, vol. 68, no. 3, pp. 249–276, 1998.

[91] B. De La Harpe, J. F. Peterson, N. Frankham, R. Zehner, D. Neale, E. Musgrave, and R. McDermott, “Assessment focus in studio: What is most prominent in architecture, art and design?,” *International Journal of Art & Design Education*, vol. 28, no. 1, pp. 37–51, 2009.

[92] A. Venables and R. Summit, “Enhancing scientific essay writing using peer assessment,” *Innovations in Education and Teaching International*, vol. 40, no. 3, pp. 281–290, 2003.

[93] N. Falchikov and J. Goldfinch, “Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks,” *Review of educational research*, vol. 70, no. 3, pp. 287–322, 2000.

[94] P. A. Carlson and F. C. Berry, “Calibrated Peer Review and assessing learning outcomes,” in *Frontiers in Education Conference*, 2003, vol. 2.

[95] R. D. Gerdeman, A. A. Russell, and K. J. Worden, “Web-Based Student Writing and Reviewing in a Large Biology Lecture Course.,” *Journal of College Science Teaching*, vol. 36, no. 5, pp. 46–52, 2007.

[96] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.

[97] C. Cheshire and J. Antin, “The social psychological effects of feedback on the production of Internet information pools,” *Journal of Computer-Mediated Communication*, vol. 13, no. 3, pp. 705–727, 2008.

[98] S. W. Huang and W. T. Fu, “Enhancing Reliability Using Peer Consistency Evaluation in Human Computation,” in *Proceedings of ACM: Computer supported collaborative work*, 2013.

[99] M. Szpir, “Clickworkers on Mars,” *American Scientist*, vol. 90, no. 3, 2002.

[100] E. H. Chi, “A Position Paper on’Living Laboratories': Rethinking Ecological Designs and Experimentation in Human-Computer Interaction,” in *Proceedings of the 13th International Conference on Human-Computer Interaction. Part I: New Trends*, 2009, pp. 597–605.

[101] S. Carter, J. Mankoff, S. R. Klemmer, and T. Matthews, “Exiting the cleanroom: On ecological validity and ubiquitous computing,” *Human--Computer Interaction*, vol. 23, no. 1, pp. 47–99, 2008.

[102] T. Lewin, *College of Future Could Be Come One, Come All*. January 2013, The New York Times, 2013.

[103] B. Efron and R. Tibshirani, *An introduction to the bootstrap*, vol. 57. Chapman & Hall/CRC, 1993.

[104] J. Ehrlinger, K. Johnson, M. Banner, D. Dunning, and J. Kruger, “Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent,” *Organizational Behavior and Human Decision Processes*, vol. 105, no. 1, pp. 98–121, 2008.

[105] S. Dow, A. Kulkarni, S. Klemmer, and B. Hartmann, “Shepherding the crowd yields better work,” in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 2012, pp. 1013–1022.

[106] D. J. Nicol and D. Macfarlane-Dick, “Formative assessment and self-regulated learning: A model and seven principles of good feedback practice,” *Studies in Higher Education*, vol. 31, no. 2, pp. 199–218, 2006.

[107] T. Gallien and J. Oomen-Early, “Personalized versus collective instructor feedback in the online courseroom: Does type of feedback affect student satisfaction, academic performance and perceived connectedness with the instructor?,” *International Journal on E-Learning*, vol. 7, no. 3, pp. 463–476, 2008.

[108] S. P. Dow, A. Glassco, J. Kass, M. Schwarz, D. L. Schwartz, and S. R. Klemmer, “Parallel prototyping leads to better design results, more divergence, and increased self-efficacy,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 17, no. 4, p. 18, 2010.

[109] J. R. Anderson and G. H. Bower, “Recognition and retrieval processes in free recall,” *Psychological review*, vol. 79, no. 2, pp. 97–123, 1972.

[110] J. Nielsen, “Enhancing the explanatory power of usability heuristics,” in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 1994, pp. 152–158.

[111] A. D. Galinsky and G. B. Moskowitz, “Counterfactuals as behavioral primes: Priming the simulation heuristic and consideration of alternatives,” *Journal of Experimental Social Psychology*, vol. 36, no. 4, pp. 384–409, 2000.

[112] J. L. Little and E. L. Bjork, “Pretesting with Multiple-choice Questions Facilitates Learning,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2012.

[113] J. Nielsen, “Iterative user-interface design,” *Computer*, vol. 26, no. 11, pp. 32–41, 1993.

[114] A. B. Markman and D. Gentner, “Splitting the differences: A structural alignment view of similarity,” *Journal of Memory and Language*, vol. 32, p. 517, 1993.

[115] N. Mazar, O. Amir, and D. Ariely, “The dishonesty of honest people: A theory of self-concept maintenance,” *Journal of marketing research*, vol. 45, no. 6, pp. 633–644, 2008.

[116] J. Kurhila, “Human-Computer Interaction by Coursera opened for credit for the students of the Department,” 2012.

[117] T. Lewin, *Students Rush to Web Classes, but Profits May Be Much Later*. January 2013, The New York Times, 2013.

[118] R. Conti, H. Coon, and T. M. Amabile, “Evidence to support the componential model of creativity: Secondary analyses of three studies,” *Creativity Research Journal*, vol. 9, no. 4, pp. 385–389, 1996.

[119] J. C. Kaufman, J. Baer, J. C. Cole, and J. D. Sexton, “A comparison of expert and nonexpert raters using the consensual assessment technique,” *Creativity Research Journal*, vol. 20, no. 2, pp. 171–178, 2008.

[120] J. E. Kuebli, R. D. Harvey, and J. H. Korn, “Critical Thinking in Critical Courses: Principles and Applications,” *Teaching Critical Thinking in Psychology: A Handbook of Best Practices*, p. 137, 2008.

[121] W. G. Perry, “Forms of intellectual development in the college years,” *New York: Holt*, 1970.

[122] L. Palen, “Social, individual and technological issues for groupware calendar systems,” in *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*, 1999, pp. 17–24.

[123] A. Fox and D. Patterson, “Crossing the software education chasm,” *Communications of the ACM*, vol. 55, no. 5, pp. 44–49, 2012.

[124] E. Roberts, J. Lilly, and B. Rollins, “Using undergraduates as teaching assistants in introductory programming courses: An update on the Stanford experience,” *ACM SIGCSE Bulletin*, vol. 27, no. 1, pp. 48–52, 1995.

[125] L. Breslow, D. Pritchard, J. DeBoer, G. Stump, A. Ho, and D. Seaton, *Studying learning in the worldwide classroom: Research into edX's first MOOC*, vol. 8. 2013, pp. 13 – 25.

[126] S. R. Klemmer, B. Hartmann, and L. Takayama, "How bodies matter: five themes for interaction design," in *Proceedings of the 6th conference on Designing Interactive systems*, 2006, pp. 140–149.

[127] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *Proceedings of the ACM SIGKDD workshop on human computation*, 2010, pp. 64–67.

[128] S. Guo, A. Parameswaran, and H. Garcia-Molina, "So who won?: dynamic max discovery with the crowd," in *Proceedings of the 2012 international conference on Management of Data*, 2012, pp. 385–396.

[129] P. Dai, M. D., and D. S. Weld, "Decision-theoretic control of crowd-sourced workflows," in *In the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*, 2010.

[130] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic Compositionality Through Recursive Matrix-Vector Spaces," in *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2012.

[131] O. F. Zaidan and C. Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, vol. 1, pp. 1220–1229.

[132] A. T. Corbett, K. R. Koedinaer, and W. Haaley, "Cognitive Tutors: From the Research Classroom to All Classrooms," *Technology enhanced learning: opportunities for change*, p. 235, 2002.

[133] R. E. Kraut and P. Resnick, *Evidence-based social design: Mining the social sciences to build online communities*. MIT Press, 2011.

[134] P. A. Murtaugh, L. D. Burns, and J. Schuster, "Predicting the retention of university students," *Research in Higher Education*, vol. 40, no. 3, pp. 355–371, 1999.

[135] J. Marlow, L. Dabbish, and J. Herbsleb, "Impression formation in online peer production: activity traces and personal profiles in github," in *Proceedings of the 2013 conference on Computer supported cooperative work*, 2013, pp. 117–128.

[136] J. J. Cadiz, A. Balachandran, E. Sanocki, A. Gupta, J. Grudin, and G. Jancke, "Distance learning through distributed collaborative video viewing," in *ACM conference on Computer supported cooperative work*, 2000, pp. 135–144.

[137] F. G. Martin, "Will massive open online courses change how we teach?," *Communications of the ACM*, vol. 55, no. 8, pp. 26–28, 2012.

[138] A. Kittur, J. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton, "The future of crowd work," in *ACM Conference on Computer Supported Coooperative Work (CSCW 2013)*, 2013.

[139] L. Hirschman, E. Breck, M. Light, J. D. Burger, and L. Ferro, "Automated grading of short-answer tests," *Intelligent Systems and their Applications, IEEE*, pp. 31–37, 2000.

[140] M. Zhang, "Contrasting Automated and Human Scoring of Essays," *R & D Connections*, 2013.

[141] H. Yannakoudakis, T. Briscoe, and B. Medlock, "A New Dataset and Method for Automatically Grading ESOL Texts.," in *ACL*, 2011, pp. 180–189.

[142] M. Chodorow and C. Leacock, "An unsupervised method for detecting grammatical errors," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, 2000, pp. 140–147.

[143] J. Burstein, M. Chodorow, and C. Leacock, "Automated essay evaluation: the criterion online writing service," *AI Magazine*, vol. 25, no. 3, p. 27, 2004.

[144] P. W. Foltz, D. Laham, and T. K. Landauer, "The intelligent essay assessor: Applications to educational technology," *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, vol. 1, no. 2, 1999.

[145] H. Chen and B. He, "Automated Essay Scoring by Maximizing Human-machine Agreement," *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

[146] M. Winerip, "Facing a Robo-Grader? Just Keep Obfuscating Mellifluously," *New York Times*, 2013.

[147] "Professionals Against Machine Scoring Of Student Essays In High-Stakes Assessment (www.humanreaders.com)."

[148] L. Hirschman, M. Light, E. Breck, and J. D. Burger, "Deep Read: A reading comprehension system," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 325–332.

[149] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Advances in neural information processing systems*, 2011, pp. 1953–1961.

[150] M. D. Peng Dai and S. Weld, "Decision-theoretic control of crowd-sourced workflows," in *In the 24th AAAI Conference on Artificial Intelligence (AAAI'10*, 2010.

[151] K. Heimerl, B. Gawalt, K. Chen, T. Parikh, and B. Hartmann, "CommunitySourcing: engaging local crowds to perform expert work via physical kiosks," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, 2012, pp. 1539–1548.

[152] J. Burstein, “The e-rater scoring engine: Automated essay scoring with natural language processing,” *Automated essay scoring: A cross-disciplinary perspective*, pp. 113–121, 2003.

[153] D. Lovallo and O. Sibony, “The case for behavioral strategy,” *McKinsey Quarterly*, pp. 30–43, 2010.

[154] D. Kahneman, D. Lovallo, and O. Sibony, “Before you make that big decision,” *Harvard Business Review*, vol. 89, no. 6, pp. 50–60, 2011.

[155] C. G. Wetzel, T. D. Wilson, and J. Kort, “The halo effect revisited: Forewarned is not forearmed.,” *Journal of Experimental Social Psychology*, 1981.

[156] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, K. Panovich, and A. Arbor, “Soylent : A Word Processor with a Crowd Inside,” *Artificial Intelligence*, pp. 313–322, 2010.

[157] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, and others, “Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey★,” *Monthly Notices of the Royal Astronomical Society*, vol. 389, no. 3, pp. 1179–1189, 2008.

[158] N. Sommers, “Revision strategies of student writers and experienced adult writers,” *College composition and communication*, vol. 31, no. 4, pp. 378–388, 1980.

[159] D. Grimes and M. Warschauer, “Utility in a fallible tool: A multi-site case study of automated writing evaluation,” *The Journal of Technology, Learning and Assessment*, vol. 8, no. 6, 2010.

[160] M. D. Shermis, C. W. Garvan, and Y. Diao, “The Impact of Automated Essay Scoring on Writing Outcomes.,” *Online Submission*, 2008.

[161] B. Yang, J.-T. Sun, T. Wang, and Z. Chen, “Effective multi-label active learning for text classification,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 917–926.

[162] E. Fast, C. Lee, A. Aiken, M. S. Bernstein, D. Koller, and E. Smith, “Crowd-scale interactive formal reasoning and analytics,” in *Proceedings of the 26th annual ACM symposium on User interface software and technology*, 2013.

[163] T. R. Guskey, “Closing Achievement Gaps: Revisiting Benjamin S. Bloom’s ‘Learning for Mastery,’” *Journal of Advanced Academics*, vol. 19, no. 1, pp. 8–31, Nov. 2007.

[164] J. A. Kulik and C.-L. C. Kulik, “Timing of Feedback and Verbal Learning.,” *Review of Educational Research*, vol. 58, no. 1, pp. 79–97, Nov. 1987.

[165] A. N. Kluger and A. DeNisi, “The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback

intervention theory.," *Psychological Bulletin*, vol. 119, no. 2, pp. 254–284, 1996.

[166] J. Hattie and H. Timperley, "The Power of Feedback," *Review of Educational Research*, vol. 77, no. 1, pp. 81–112, Mar. 2007.

[167] F. E. Balcazar, B. L. Hopkins, and Y. Suarez, "A critical, objective review of performance feedback.," *Journal of Organizational Behavior Management*, vol. 7, no. 2, 1986.

[168] G. P. Latham and E. A. Locke, "Self-regulation through goal setting," *Organizational Behavior and Human Decision Processes*, vol. 50, no. 2, pp. 212–247, Dec. 1991.

[169] N. Heffernan, C. Heffernan, K. Dietz, D. Soffer, S. R. Pellegrino, J. W. Goldman, and M. Dailey, "Improving Mathematical Learning Outcomes Through Automatic Reassessment and Relearning," in *AERA*, 2012.

[170] S. Anderson and J. Rodin, "Is Bad News Always Bad?: Cue and Feedback Effects on Intrinsic Motivation," *Journal of Applied Social Psychology*, vol. 19, no. 6, pp. 449–467, May 1989.

[171] R. M. Dawes, "The robust beauty of improper linear models in decision making.," *American psychologist*, vol. 34, no. 7, p. 571, 1979.

[172] C. Kulkarni, R. Socher, M. S. Bernstein, and S. R. Klemmer, "Scaling Short-answer Grading by Combining Peer Assessment with Algorithmic Scoring," in *ACM Conf on Learning@Scale*, 2014.

[173] P. André, M. Bernstein, and K. Luther, "Who gives a tweet?," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*, 2012, p. 471.

[174] R. L. Marsh, J. D. Landau, and J. L. Hicks, "How examples may (and may not) constrain creativity.," *Memory & cognition*, vol. 24, no. 5, pp. 669–80, Sep. 1996.

[175] W. D. Gray and D. A. Boehm-Davis, "Milliseconds matter: an introduction to microstrategies and to their use in describing and predicting interactive behavior.," *Journal of experimental psychology. Applied*, vol. 6, no. 4, pp. 322–35, Dec. 2000.

[176] J. A. Krosnick, "Survey research," *Annual review of psychology*, vol. 50, no. 1, pp. 537–567, 1999.

[177] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006, pp. 423–430.

[178] N. Sommers, "Responding to Student Writing," *College Composition and Communication*, vol. 33, no. 2, pp. 148–156, 1982.

[179] S. Dow, J. Fortuna, D. Schwartz, B. Altringer, and S. Klemmer, “Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results,” in *Proceedings of the 2011 annual conference on Human factors in computing systems*, 2011, pp. 2807–2816.

[180] S. P. Dow, K. Heddleston, and S. R. Klemmer, “The efficacy of prototyping under time constraints,” in *Proceeding of the ACM conference on Creativity and cognition*, 2009, p. 165.

[181] A. Kittur, E. H. Chi, and B. Suh, “Crowdsourcing user studies with Mechanical Turk,” in *Proc. of CHI*, 2008, p. 453.

[182] C. Kulkarni, J. Cambre, Y. Kotturi, M. S. Bernstein, and S. Klemmer, “Talkabout: Making Distance Matter with Small Groups in Massive Classes,” in *Proceedings of CSCW 2015: ACM Conference on Computer Supported Collaborative Work*, 2015.

[183] C. J. Nemeth, “Differential contributions of majority and minority influence.,” *Psychological Review*, vol. 93, no. 1. 1986.

[184] J. Kucsera and G. Orfield, “New York State’s Extreme School Segregation: Inequality, Inaction and a Damaged Future,” 2014.

[185] K. Olds, “Mapping Coursera’s Global Footprint,” *Inside Higher Ed*, 2013. [Online]. Available: http://www.insidehighered.com/blogs/globalhighered/mapping-courseras-global-footprint.

[186] J. A. Konstan, J. D. Walker, D. C. Brooks, K. Brown, and M. D. Ekstrand, “Teaching recommender systems at large scale,” in *Proc of the ACM conference on Learning @ scale conference*, 2014.

[187] A. J. Jacobs, “Grading the MOOC University,” *The New York Times*, 20-Apr-2013.

[188] E. Losh, *The War on Learning: Gaining Ground in the Digital University*. MIT Press, 2014.

[189] E. G. T. Green, “Variation of Individualism and Collectivism within and between 20 Countries: A Typological Analysis,” *Journal of Cross-Cultural Psychology*, vol. 36, no. 3, pp. 321–339, May 2005.

[190] B. Goesling, “Changing Income Inequalities within and between Nations: New Evidence,” *American Sociological Review*, vol. 66, no. 5, pp. 745–761, 2001.

[191] N. Purdie and J. Hattie, “Cultural Differences in the Use of Strategies for Self-Regulated Learning,” *American Educational Research Journal*, vol. 33, no. 4, pp. 845–871, Jan. 1996.

[192] X. Lin and D. L. Schwartz, “Reflection at the Crossroads of Cultures,” *Mind, Culture, and Activity*, vol. 10, no. 1, Feb. 2003.

[193] J. Freedman, “MOOCs Are Usefully Middlebrow,” *The Chronicle of Higher Education*.

[194] R. L. Coser, “The Complexity of Roles as a Seedbed of Individual Autonomy,” in *The Idea of Social Structure: Essays in Honor of Robert Merton*, New York, New York, USA, 1975.

[195] L. A. Braskamp, D. C. Braskamp, and K. Merrill, “Assessing Progress in Global Learning and Development of Students with Education Abroad Experiences.,” *Frontiers: The Interdisciplinary Journal of Study Abroad*, vol. 18, pp. 101–118, Nov. 2008.

[196] J. Tudge, *The everyday lives of young children*. Cambridge, UK: Cambridge University Press, 2008.

[197] S. J. Heine, *Cultural Psychology*. WW Norton, 2008.

[198] H. Markus and S. Kitayama, “Culture and the self: Implications for cognition, emotion, and motivation,” *Psychological Review*, vol. 98, no. 2, pp. 224–253, 1991.

[199] I. Varnum, Michael EW Grossmann, S. Kitayama, and R. E. Nisbett, “The origin of cultural differences in cognition the social orientation hypothesis,” *Current directions in psychological science*, vol. 19, no. 1, 2010.

[200] E. Rocco, “Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact,” in *Proc of CHI: ACM Conf on Human Factors in Computing Systems*, 1998.

[201] J. E. Short, E. Williams, and B. Christie, *The Social Psychology of Telecommunications*. London: Wiley, 1976.

[202] A. J. Saltarelli, “Effects of belongingness and synchronicity on face-to-face and computer-mediated online cooperative pedagogy,” Jan. 2012.

[203] R. L. Daft and R. H. Lengel, “Organizational Information Requirements, Media Richness and Structural Design,” *Management Science*, vol. 32, no. 5, pp. 554–571, May 1986.

[204] G. DeSanctis, A.-L. Fayard, M. Roach, and L. Jiang, “Learning in Online Forums,” *European Management Journal*, vol. 21, no. 5, pp. 565–577, Oct. 2003.

[205] D. Coetzee, A. Fox, M. A. Hearst, and B. Hartmann, “Chatrooms in MOOCs: all talk and no action,” in *Proc. of the ACM conference on Learning @ scale*, 2014, pp. 127–136.

[206] K. Papadopoulos, L. Sritanyaratana, and S. R. Klemmer, “Community TAs scale high-touch learning, provide student-staff brokering, and build esprit de corps,” in *Proceedings of the first ACM conference on Learning @ scale conference - L@S ’14*, 2014, pp. 163–164.

[207] W. B. Group, *World Development Indicators 2012*. World Bank Publications, 2012.

[208] A. M. O'Donnell and D. F. Dansereau, "Scripted cooperation in student dyads: A method for analyzing and enhancing academic learning and performance," in *Interaction in cooperative groups: The theoretical anatomy of group learning*, Cambridge, UK: Cambridge University Press, 1995, pp. 120–141.

[209] M. Nguyen, Y. S. Bin, and A. Campbell, "Comparing online and offline self-disclosure: a systematic review.," *Cyberpsychology, behavior and social networking*, vol. 15, no. 2, Mar. 2012.

[210] A. N. Joinson, "Knowing Me, Knowing You: Reciprocal Self-Disclosure in Internet-Based Surveys," *CyberPsychology & Behavior*, vol. 4, no. 5, Oct. 2001.

[211] B. J. Guzzetti and A. Others, "Promoting Conceptual Change in Science: A Comparative Meta-Analysis of Instructional Interventions from Reading Education and Science Education.," *Reading Research Quarterly*, vol. 28, no. 2, pp. 116–59, Nov. 1992.

[212] S. L. Star and J. R. Griesemer, "Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39," *Social Studies of Science*, vol. 19, no. 3, pp. 387–420, Aug. 1989.

[213] W. C. Parker, "Classroom Discussion: Models for Leading Seminars and Deliberations.," *Social Education*, vol. 65, no. 2, pp. 111–15, Nov. 2000.

[214] S. D. Brookfield and S. Preskill, *Discussion as a Way of Teaching: Tools and Techniques for Democratic Classrooms*. John Wiley & Sons, 2012.

[215] A. D. Ho, J. Reich, S. O. Nesterko, D. T. Seaton, T. Mullaney, J. Waldo, and I. Chuang, "HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013," *SSRN Electronic Journal*, Jan. 2014.

[216] J. H. Tomkin and D. Charlevoix, "Do professors matter?," in *Proc. of the ACM conference on Learning @ scale*, 2014, pp. 71–78.

[217] M. Desai, "Human development: Concepts and measurement," *European Economic Review*, vol. 35, no. 2–3, pp. 350–357, Apr. 1991.

[218] S. Shenkar and R. Oded, "Clustering Countries on Attitudinal Dimensions: A Review and Synthesis," *The Academy of Management Review*, vol. 10, no. 3, pp. 435–454, 1985.

[219] M. Marmot, "Social determinants of health inequalities.," *Lancet*, vol. 365, no. 9464, pp. 1099–104, Jan. 2005.

[220] *World Values Survey (Wave 6)*. World Values Survey Association, 2014.

[221] F. Clementi and Mauro Gallegati, “Econophysics of Wealth Distributions,” in *Econophysics of Wealth Distributions*, A. Chatterjee, S. Yarlagadda, and B. K. Chakrabarti, Eds. Milano: Springer Milan, 2005.

[222] D. C. Hambrick, S. C. Davison, S. A. Snell, and C. C. Snow, “When Groups Consist of Multiple Nationalities: Towards a New Understanding of the Implications,” *Organization Studies*, vol. 19, no. 2, pp. 181–205, Mar. 1998.

[223] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone, “Evidence for a collective intelligence factor in the performance of human groups.,” *Science (New York, N.Y.)*, vol. 330, no. 6004, pp. 686–8, Oct. 2010.

[224] E. Mazur, “Farewell, lecture?,” *Science (New York, N.Y.)*, vol. 323, no. 5910, pp. 50–1, Jan. 2009.

[225] C. H. Crouch and E. Mazur, “Peer Instruction: Ten years of experience and results,” *American Journal of Physics*, vol. 69, no. 9, p. 970, Sep. 2001.

[226] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, “Engaging with massive online courses,” pp. 687–698, Apr. 2014.

[227] A. Nora and A. F. Cabrera, “The Role of Perceptions in Prejudice and Discrimination and the Adjustment of Minority Students to College.,” *Journal of Higher Education*, vol. 67, no. 2, pp. 119–48, Nov. 1995.

[228] S. Hurtado, D. F. Carter, and D. Kardia, “The Climate for Diversity: Key Issues for Institutional Self-Study,” *New Directions for Institutional Research*, vol. 1998, no. 98, pp. 53–63, 1998.

[229] J. U. Ogbu, “Understanding Cultural Diversity and Learning,” *Educational Researcher*, vol. 21, no. 8, pp. 5–14, Nov. 1992.

[230] G. Mark, J. Grudin, and S. E. Poltrock, “Meeting at the Desktop: An Empirical Study of Virtually Collocated Teams,” in *ECSCW*, 1999.

[231] P. Dillenbourg, “Over-scripting CSCL: The risks of blending collaborative learning with instructional design.,” *Three worlds of CSCL. Can we support CSCL?*, pp. 61–91, 2002.

[232] N. Bos, J. Olson, D. Gergle, G. Olson, and Z. Wright, “Effects of four computer-mediated communications channels on trust development,” in *Proc CHI: ACM conference on Human factors in computing systems*, 2002.

[233] S. Turkle, *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books, 2011.

[234] C. Kulkarni, J. Cambre, Y. Kotturi, M. Bernstein, and S. Klemmer, “Talkabout: Making distance matter with small groups in massive classes,” *CSCW: ACM Conference on Computer Supported Collaborative Work*, 2015.

[235] L. Porter, C. B. Lee, and B. Simon, “Halving fail rates using peer instruction: a study of four computer science courses,” *Proceeding SIGCSE ’13 Proceeding*

of the 44th ACM technical symposium on Computer science education, pp. 177–182, 2013.

[236] M. Smith, W. Wood, W. Adams, C. Wieman, J. Knight, N. Guild, and T. Su, "Why Peer Discussion Improves Student Performance on In-Class Concept Questions," *Science*, vol. 323, no. 5910, pp. 122–124, 2009.

[237] J. Bransford and D. L. Schwartz, "Rethinking Transfer: A Simple Proposal with Multiple Implications," *Review of Research in Education*, vol. 24, pp. 61–100, 1999.

[238] J. Bransford, A. Brown, and R. Cocking, *How People Learn*. 2000.

[239] J. Cambre, C. Kulkarni, M. S. Bernstein, and S. R. Klemmer, "Talkabout: Small-group Discussions in Massive Global Classes," in *Learning@Scale*, 2014.

[240] P. Dourish and G. Bell, "The infrastructure of experience and the experience of infrastructure: meaning and structure in everyday encounters with space," *Environment and Planning B Planning and Design*, vol. 34, no. 3, p. 414, 2007.

[241] T. Erickson and W. Kellogg, "Social translucence: an approach to designing systems that support social processes," *ACM Transactions on Computer-Human Interaction (TOCHI) - Special issue on human-computer interaction in the new millennium, Part 1*, vol. 7, no. 1, pp. 59–83, 2000.

[242] P. Dourish and V. Bellotti, "Awareness and coordination in shared workspaces," *Proceeding CSCW '92 Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pp. 107–114, 1992.

[243] K. Stephens-Martinez, M. A. Hearst, and A. Fox, "Monitoring MOOCs: which information sources do instructors value?," *Proceedings of the first ACM conference on Learning @ scale conference*, pp. 79–88, 2014.

[244] J. Rosenberg, M. Lorenzo, and E. Mazur, "Peer instruction: Making science engaging," *Handbook of college science teaching*, pp. 77–85, 2006.

[245] R. E. Kraut, P. Resnick, S. Kiesler, M. Burke, Y. Chen, N. Kittur, J. Konstan, Y. Ren, and J. Riedl, *Building Successful Online Communities: Evidence-Based Social Design (Google eBook)*. MIT Press, 2012.

[246] J. Grudin, "Groupware and social dynamics: eight challenges for developers," *Communications of the ACM*, vol. 37, no. 1, pp. 92–105, 1994.

[247] R. F. Kizilcec and E. Schneider, "Motivation as a Lens to Understand Online Learners: Towards Data-Driven Design with the OLEI Scale."

[248] D. Coetzee, A. Fox, M. A. Hearst, and B. Hartmann, "Chatrooms in MOOCs: all talk and no action," in *Proc. of the ACM conference on Learning @ scale*, 2014, pp. 127–136.

[249] D. Coetzee, S. Lim, A. Fox, B. Hartmann, and M. A. Hearst, “Structuring Interactions for Large-Scale Synchronous Peer Learning,” *CSCW: ACM Conference on Computer Supported Collaborative Work*, 2015.

[250] R. Cialdini and N. Goldstein, “Social influence: Compliance and conformity,” *Annual review of psychology*, vol. 55, pp. 591–621, 2004.

[251] K. Ling and G. Beenen, “Using Social Psychology to Motivate Contributions to Online Communities,” *Journal of Computer-Mediated Communication*, vol. 10, no. 4, p. 00, 2005.

[252] J. Cheng, L. Adamic, P. Dow, J. Kleinberg, and J. Leskovec, “Can cascades be predicted?,” *Proceedings of the 23rd international conference on World wide web*, pp. 925–936, 2014.

[253] E. Bakshy, B. Karrer, and L. A. Adamic, “Social influence and the diffusion of user-created content,” *Proceedings of the 10th ACM conference on Electronic commerce*, pp. 325–334, 2009.

[254] M. Chen, “Design of a virtual auditorium,” *MULTIMEDIA ’01 Proceedings of the ninth ACM international conference on Multimedia*, pp. 19–28, 2001.

[255] S. R. Klemmer, *Katayanagi Lecture at CMU: The Power of Examples*. p. 2011.

[256] H. Kim and P. Hinds, “Harmony vs. disruption: The effect of iterative prototyping on teams’ creative processes and outcomes in the West and the East,” in *Proc. ICIC: International Conf. on Intercultural Collaboration*, 2012.

[257] A. Ritter, S. Clark, O. Etzioni, and others, “Named entity recognition in tweets: an experimental study,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1524–1534.

[258] T. Piketty, “Capital in the 21st Century,” *Cambridge: Harvard Uni*, 2014.

[259] T. Leventhal and J. Brooks-Gunn, “The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes.,” *Psychological bulletin*, vol. 126, no. 2, p. 309, 2000.

[260] T. Leventhal and J. Brooks-Gunn, “Moving to opportunity: an experimental study of neighborhood effects on mental health,” *American Journal of Public Health*, vol. 93, no. 9, pp. 1576–1582, 2003.

[261] G. Der, S. Macintyre, G. Ford, K. Hunt, and P. West, “The relationship of household income to a range of health measures in three age cohorts from the West of Scotland,” *The European Journal of Public Health*, vol. 9, no. 4, pp. 271–277, 1999.

[262] I. Weber and C. Castillo, “The demographics of web search,” in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, pp. 523–530.

[263] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli, “The singularity is not near: slowing growth of Wikipedia,” in *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, 2009, p. 8.

[264] M. Burke, C. Marlow, and T. Lento, “Feed me: motivating newcomer contribution in social network sites,” in *Proceedings of the 27th international conference on Human factors in computing systems*, 2009, pp. 945–954.

[265] M. Burke and B. Settles, “Plugged in to the community: Social motivators in online goal-setting groups,” in *Proceedings of the 5th International Conference on Communities and Technologies*, 2011, pp. 1–10.

[266] O. Nov, M. Naaman, and C. Ye, “Motivational, structural and tenure factors that impact online community photo sharing,” 2009.

[267] E. Goffman, “The presentation of self in everyday life. 1959,” *Garden City, NY*, 2002.

[268] M. Sherif, “A study of some social factors in perception.,” *Archives of Psychology (Columbia University)*, 1935.

[269] C. Kulkarni and E. Chi, “All the news that’s fit to read: a study of social annotations for news reading,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2407–2416.

[270] C. E. Kasworm, “An examination of self-directed contract learning as an instructional strategy,” *Innovative Higher Education*, vol. 8, no. 1, pp. 45–54, 1983.

[271] M. J. Salganik, P. S. Dodds, and D. J. Watts, “Experimental study of inequality and unpredictability in an artificial cultural market,” *science*, vol. 311, no. 5762, pp. 854–856, 2006.

[272] M. Speca, L. E. Carlson, E. Goodey, and M. Angen, “A randomized, wait-list controlled clinical trial: the effect of a mindfulness meditation-based stress reduction program on mood and symptoms of stress in cancer outpatients,” *Psychosomatic medicine*, vol. 62, no. 5, pp. 613–622, 2000.

[273] B. S. Frey and R. Jegen, “Motivation crowding theory: A survey of empirical evidence,” 2000.

[274] D. Gentner, “Structure-mapping: A theoretical framework for analogy,” *Cognitive science*, vol. 7, no. 2, pp. 155–170, 1983.

[275] D. L. Schwartz, C. C. Chase, and J. D. Bransford, “Resisting overzealous transfer: Coordinating previously successful routines with needs for new learning,” *Educational Psychologist*, vol. 47, no. 3, pp. 204–214, 2012.

[276] J. G. Politz, D. Patterson, S. Krishnamurthi, and K. Fisler, “CaptainTeach: Multi-stage, in-flow peer review for programming assignments,” in *Proceedings of the 2014 conference on Innovation & technology in computer science education*, 2014, pp. 267–272.

[277] Y. Lu, J. Warren, C. Jermaine, S. Chaudhuri, and S. Rixner, “Grading the Graders: Motivating Peer Graders in a MOOC,” in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 680–690.

[278] D. Prelec, “A Bayesian truth serum for subjective data,” *science*, vol. 306, no. 5695, pp. 462–466, 2004.

[279] P. J. Hinds, “The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance.,” *Journal of Experimental Psychology: Applied*, vol. 5, no. 2, p. 205, 1999.

[280] C. M. Hicks, C. A. Fraser, P. Desai, and S. Klemmer, “Do Numeric Ratings Impact Peer Reviewers?,” in *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 2015, pp. 359–362.

[281] B. M. Bass and B. J. Avolio, “Potential biases in leadership measures: How prototypes, leniency, and general satisfaction relate to ratings and rankings of transformational and transactional leadership constructs,” *Educational and psychological measurement*, vol. 49, no. 3, pp. 509–527, 1989.

[282] K. Raman and T. Joachims, “Methods for ordinal peer grading,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1037–1046.

[283] A. E. Waters, D. Tinapple, and R. G. Baraniuk, “BayesRank: A Bayesian Approach to Ranked Peer Grading,” in *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 2015, pp. 177–183.

[284] H. Zhu, S. P. Dow, R. E. Kraut, and A. Kittur, “Reviewing versus doing: Learning and performance in crowd assessment,” in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014, pp. 1445–1455.

[285] J. Tang, J. Marlow, A. Hoff, A. Roseway, K. Inkpen, C. Zhao, and X. Cao, “Time travel proxy: using lightweight video recordings to create asynchronous, interactive meetings,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 3111–3120.

[286] M. L. Stevens, E. A. Armstrong, and R. Arum, “Sieve, incubator, temple, hub: Empirical and theoretical advances in the sociology of higher education,” *Annu. Rev. Sociol*, vol. 34, pp. 127–151, 2008.

[287] E. Eide, D. J. Brewer, and R. G. Ehrenberg, “Does it pay to attend an elite private college? Evidence on the effects of undergraduate college quality on graduate school attendance,” *Economics of Education Review*, vol. 17, no. 4, pp. 371–376, 1998.

[288] B. DiSalvo, C. Reid, and P. K. Roshan, “They can’t find us: the search for informal CS education,” in *Proceedings of the 45th ACM technical symposium on Computer science education*, 2014, pp. 487–492.

[289] C. T. Clotfelter, R. G. Ehrenberg, M. Getz, and J. J. Siegfried, *Economic challenges in higher education*. University of Chicago Press, 2008.

[290] D. L. Schwartz and D. Arena, *Measuring What Matters Most*. The MIT Press, 2013.

[291] C. E. Kulkarni, M. S. Bernstein, and S. R. Klemmer, “PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance,” in *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 2015, pp. 75–84.

[292] Y. Kotturi, C. Kulkarni, M. Bernstein, and S. Klemmer, “Structure and messaging techniques for online peer learning systems that increase stickiness,” in *Learning at Scale (to appear)*, 2015.

[293] “Digest of Educational Statistics,” 2013.