Peer and Self Assessment in Massive Online Classes

CHINMAY KULKARNI[1], KOH PANG WEI[1,2], HUY LE[2], DANIEL CHIA[1,2], KATHRYN PAPADOPOULOS[1], JUSTIN CHENG[1], DAPHNE KOLLER[1,2], SCOTT R. KLEMMER[1]
[1]Stanford University; [2]Coursera, Inc.

Peer and self assessment offer an opportunity to scale both assessment and learning to global classrooms. This paper reports our experiences with two iterations of the first large online class to use peer and self assessment. In this class, peer grades correlated highly with staff-assigned grades. The second iteration had 42.9% of students grades within 5% of the staff grade, and 65.5% within 10%. On average, students assessed their work 7% higher than staff did. Students also rated peers' work from their own country 3.6% higher than those from elsewhere. We performed three experiments to improve grading accuracy. We found that giving students feedback about their grading bias increased subsequent accuracy. We introduce short, customizable feedback snippets that cover common issues with assignments, providing students more qualitative peer feedback. Finally, we introduce a data-driven approach that highlights high-variance items for improvement. We find that rubrics that use a parallel sentence structure, unambiguous wording and well-specified dimensions have lower variance. After revising rubrics, median grading error decreased from 12.4% to 9.9%.

Categories and Subject Descriptors: K.3.1 [**Computer Uses in Education**]: Distance learning, Collaborative learning; H.5.3 [**Group and Organization Interfaces**]: Computer-supported cooperative work

General Terms: Design, Experimentation, Human Factors, Performance

Additional Key Words and Phrases: Peer assessment, self assessment, MOOC, online education, Massive online classroom, design assessment, qualitative feedback, design crit, studio-based learning

ACM Reference Format:
Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, Scott R. Klemmer, 2013. Peer and Self Assessment in Massive Online Classes. *ACM Trans. Comput.-Hum. Interact.* 9, 4, Article 39 (March 2013), 31 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

In the past year, hundreds of thousands of students have earned certificates in large online classes—on topics from Databases to Sociology to World Music—and millions have signed up [Lewin 2012a]. These classes, often called 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. [Widom 2012]), which precludes the open-ended work that is a hallmark of education in creative fields like design [Buxton 2007]. Furthermore, viewing and critiquing others' work plays a key pedagogical role in these domains [Schön 1985]. Fields like design have also traditionally relied on intimate co-location to enable these activities and to confer values and norms [Schön

This work is supported by the Hasso Plattner Design Thinking Program and NSF CAREER award IIS-0745320.

DOI:http://dx.doi.org/10.1145/0000000.0000000

1985]. However, in a global, online classroom, students lack the shared context colocation 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 [Hearst 2000]. But, do students have the motivation and expertise to perform peer assessment well? This paper reports on our experiences with the first use of peer assessment in a massive online class. It is the largest use of peer assessment to date. As of June 2013, this technique has since been adopted in many other classes, including 79 MOOCs on the Coursera[1] platform alone.

1.1. 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 [Lawson 2006], HCI [Winograd 1990; Greenberg 2009], and software design [Tomayko 1991]. This paper considers the studio as an inspiration for online design education.

The studio model of education was formalized in the École de Beaux-Arts [Drexler et al. 1977]. Studios provide an open, shared environment for students to work. This copresence provides social motivation and facilitates peer learning through visibility of work [Reimer and Douglas 2003]. Formal and informal studio critique helps students iteratively improve their work [Schön 1985].

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 [Tinapple et al. 2013].

Formative studio feedback further engages students in reflective practice [Schön 1985]. Informal, formative feedback is often through oral critiques or "crits" by teachers or other experts [Uluoglu 2000]. Such informal, qualitative feedback is essential, because it encourages iterative practice [Cennamo et al. 2011]. Because crits are often delivered in public, students also learn from observing peer work as well as by working on their own [Dannels and Martin 2008].

Expert critiques also serve as summative assessment. Experts often assess design based on trained but tacit criteria [Snodgrass and Coyne 2006]. Amabile *et al* demonstrate that expert consensus is a reliable measure of the quality of creative work [Amabile 1982]. Their Consensual Assessment 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 [Feldman 1994].

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 [Tohidi et al. 2006; Buxton 2007], and on exploration and iteration [Fallman 2003]. Second, assessment must allow students to learn the tacit criteria of good design. Criteria for good design are often not explicitly defined [Forlizzi and Battarbee 2004]. For instance, interactive interfaces may be subjectively evaluated for whether they are learnable and appropriate [Alben

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

1996], criteria that require tacit interpretation. Third, assessment must provide students both qualitative formative feedback, and summative feedback.

1.2. The promise of peer assessment

The inherent variability of open-ended solutions, and lack of defined evaluation criteria for design makes automatically assessing open-ended work challenging [Bennett et al. 1997]. In addition, automated systems frequently cannot capture the semantic meaning of answers, which limits the feedback that they can provide to help students improve [Bennett 1998; Hearst 2000].

Therefore, open-ended assignments generally rely on human graders. The time-intensive, personalized assessment of grading sketches, designs, and other open-ended assignments requires a small student-to-grader ratio [Hsi and Agogino 1995; Stanley and Porter 2002]. This staff effort is prohibitive for large classes: staff grading simply doesn't scale.

Peer and self assessment is a promising alternative, with potential additional benefits. 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 [Tinapple et al. 2013]. Evaluating peers' work also exposes students to solutions, strategies, and insights that they otherwise would likely not see [Chinn 2005; Tinapple et al. 2013]. Similarly, self assessment helps students reflect on gaps in their understanding, making them more resourceful, confident, and higher achievers [Zimmerman and Schunk 2001; Pintrich 1995; Pintrich and Zusho 2007] and provides learning gains not seen with external evaluation [Dow et al. 2012].

Peer assessment can increase student involvement and maturity, lower the grading burden on staff, and enhance classroom discussion [Boud 1995]. Peer assessment has been used in colocated classroom settings for many different kinds of assignments [Topping 1998], including design [De La Harpe et al. 2009; Tinapple et al. 2013]. programming [Chinn 2005] and essays [Venables and Summit 2003]. How can we make this classroom technique scale to a large online class?

1.3. Scaling peer assessment

In-class peers can assess each other well [Falchikov and Goldfinch 2000; Carlson and Berry 2003; Gerdeman et al. 2007]. To effectively scale peer assessment, we can learn several lessons from crowdsourcing [Surowiecki 2005]. First, crowdworkers perform better when they are intrinsically motivated by the task's importance [Cheshire and Antin 2008]. Second, consensus among raters serves as a useful indicator of quality [Huang and Fu 2013]. Third, interfaces like FoldIt [Khatib et al. 2011] and NASA Clickworkers [Szpir 2002] demonstrate that short, well-crafted training exercises can enable legions of motivated amateurs to perform work previously thought to require years of training.

Massive online classes provide a valuable living lab [Chi 2009; Carter et al. 2008] for exploring peer-sourcing approaches, and our hope is that peer-sourcing insights from massive classes will contribute techniques that apply more broadly. These peer-sourced systems introduce new challenges and opportunities beyond crowdsourcing. For example, students using peer assessment both create the work to be assessed *and* perform the assessment. One theme this paper will explore is the learning benefits that arise from those dual roles.

1.4. Contributions

This paper 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%.

2. THE ANATOMY OF A LARGE SCALE 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 1).

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.

2.1. By the numbers

Similar to other online classes [Lewin 2013a], 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 2). 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.

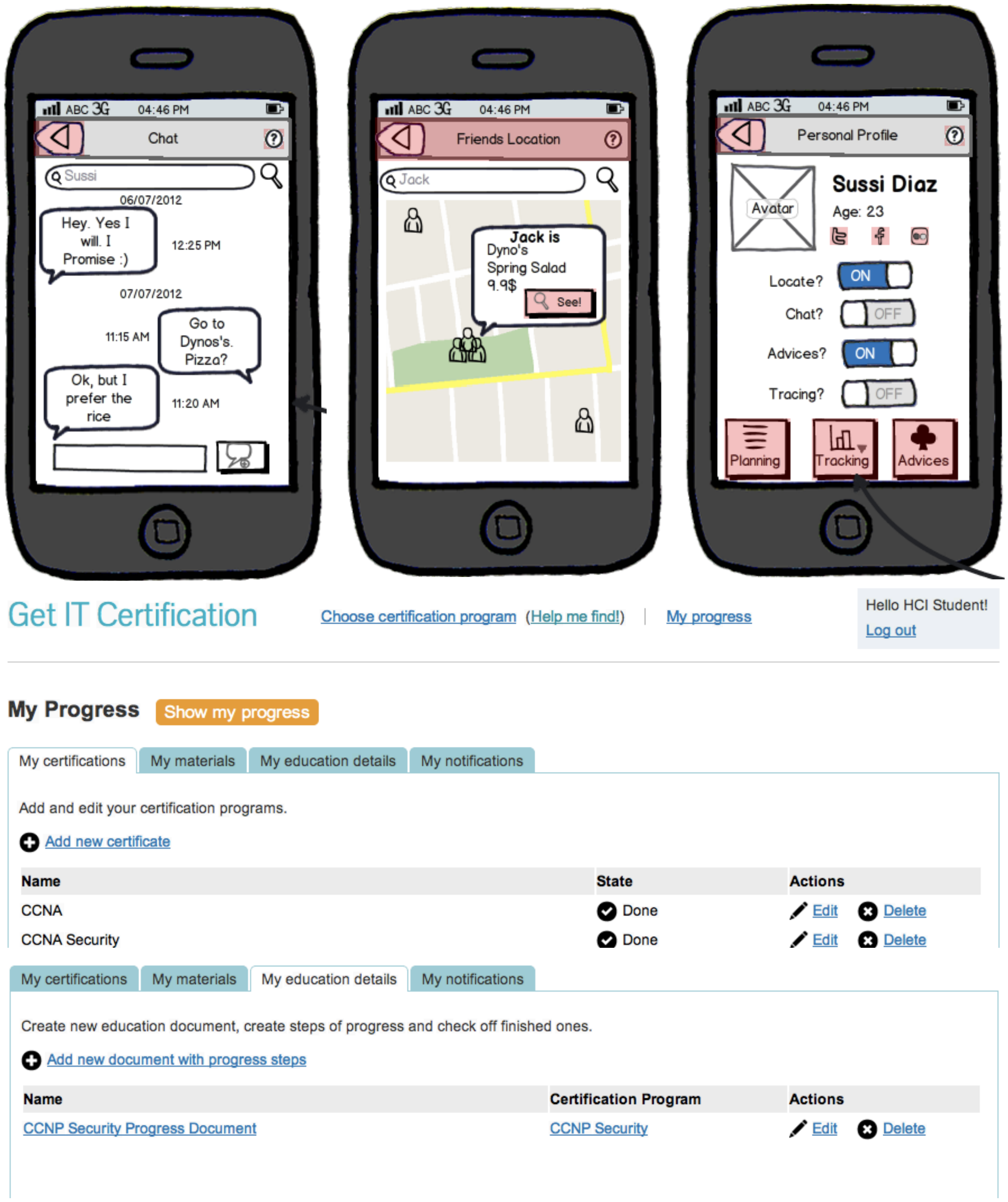


Fig. 1. Example prototypes from student projects in the online class (top: early prototype of a social dining app; bottom: a tracker for professional certification at the end of class).

In all, 2,673 students submitted assignments in the first iteration, and 3,203 in the second (Figure 3). 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.

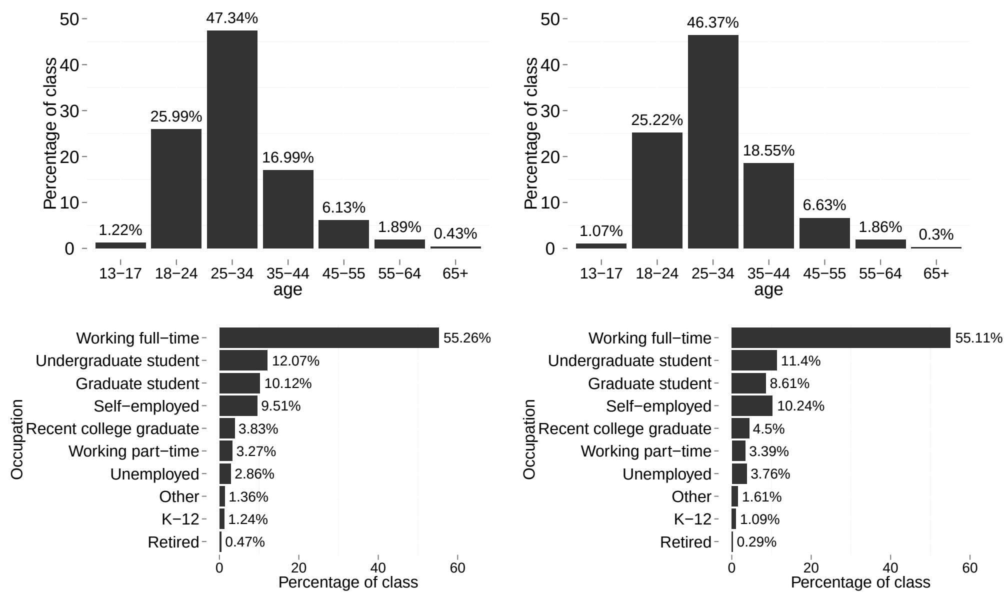
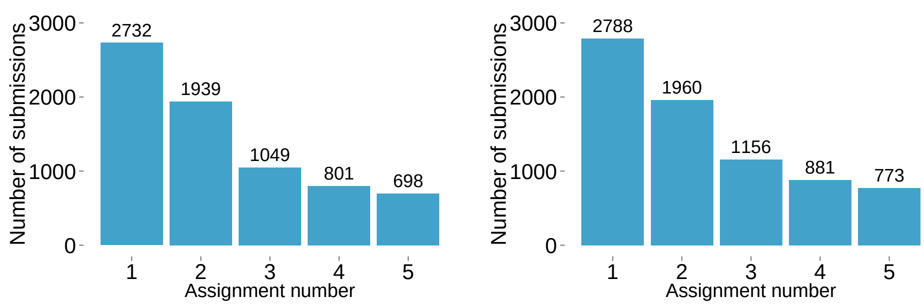


Fig. 2. Online classes attract students who cannot use traditional universities, such as those working full-time. The age distribution of the class is remarkably similar across both iterations. (a) Spring 2012 (iteration 1), 10190 participants, (b) Fall 2012 (iteration 2), 17915 participants.



(a) Spring 2012 (iteration 1)

(b) Fall 2012 (iteration 2)

Fig. 3. Number of students who submitted each assignment.

2.2. Assignments

All assignments were submitted online, and graded with calibrated peer assessment. Some 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 [Andrade 2005]. 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[2].

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[3]. The in-person Stanford class uses self assessment and staff grading, but not peer assessment.

2.3. Peer Assessment

Assessment used Calibrated Peer Review [Carlson and Berry 2003]. Calibrated peer review helps students learn to grade by first practicing grading on sample submissions.

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[4].

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

[2]All assessment materials are also available in full at http://hci.st/assess
[3]https://cs147.stanford.edu/
[4]Times for the lower 75% of submissions provide an approximate upper bound to the grading burden. We use the lower 75% to exclude assessments that weren't completed or ones completed over multiple log-in sessions.

Table I.

Guiding questions	Bare minimum	Satisfactory effort & performance	Above & Beyond
		. . .	
Alternate redesign—Extra credit. Have you created a fully functional alternate prototype?	0: No URL to functional prototype	3: URL present, but prototype only partially functional.	5: URL present, Alternative prototype is complete.
User testing. Photographs—extra credit. Did you submit photos from all three user testing sessions?	0: No photographs were uploaded.	3: Some photographs were uploaded (but less than 3), OR photos don't show an interesting moment in the experiment (e.g. photograph of participant signing consent form is not an interesting photo).	5: At least 3 photographs are uploaded and all photographs show interesting moments in the evaluation. Photos have meaningful captions
		. . .	

Category	Unsatisfactory	Bare minimum	Satisfactory effort & performance	Above & Beyond
		. . .		
Extra Credit: Electronic Prototype of Redesign	0: No URL to functional prototype	1: The prototype is incomplete and barely interactive.	3: The prototype is somewhat interactive, but not ready for user testing.	5: The alternative prototype is fully interactive and ready for user testing.
Photos/Sketches	0: No photographs were submitted that showed interesting moments in the user testing process.	1: 1 photograph was submitted that showed an interesting moment in the user testing process.	3: 2 photographs were submitted that showed interesting moments in the user testing process..	5: 3 or more photographs were submitted that showed interesting moments in the user testing process.
		. . .		

Above A fragment of the original rubric for the last assignment. Only two of six questions are shown, the rest are above and below these (shown as ellipses)
Below Fragment of revised rubric for the same questions. The new rubric uses *categories* instead of guiding questions, introduces a new column for completely missing and unsatisfactory work, and uses a parallel sentence structure.

(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.

3. HOW ACCURATE WAS PEER ASSESSMENT?

3.1. 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 (these were randomly selected). 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 paper'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 [Efron and Tibshirani 1993]. 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.

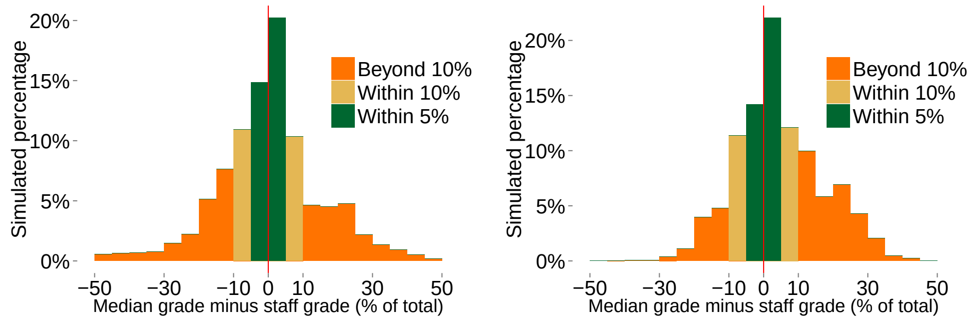
At the end of the class, students were invited to participate in a survey; 3,550 students participated in all. Participation was voluntary, students were not compensated, and the survey did not count towards course credit.

3.2. Results: Grading agreement

Here, we present percentage differences between peer and staff grades (summarized in Table II). 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).

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 4). The second iteration improved to 42.9% within 5% of the staff grade, and 65.5% within 10%. In the first iteration of

[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.



(a) Iteration 1: 34.0% of samples within 5% of the staff grade, and 56.9% within 10%.

(b) Iteration 2:. 42.0% of samples within 5% of the staff grade, and 65% within 10%.

Fig. 4. 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.

Table II. 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%

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).

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 5). The median submission had a self grade 6% higher than the median peer grade. In the second iteration, 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.)

3.3. 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%. 28% of submissions 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%.

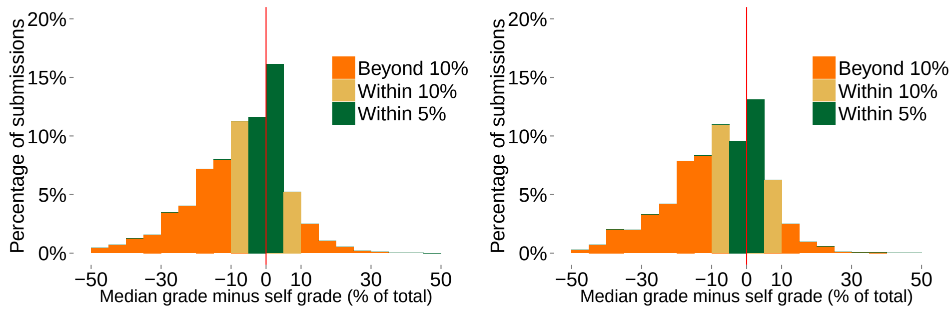


Fig. 5. (a) Comparison of median peer grades against self grades. In the first iteration 28.7% of such samples were within 5% of the staff grade, and 44.9% within 10%. (b) Same graph for second iteration of the class. 24.0% of such samples were within 5% of the staff grade, and 40.63% within 10%.

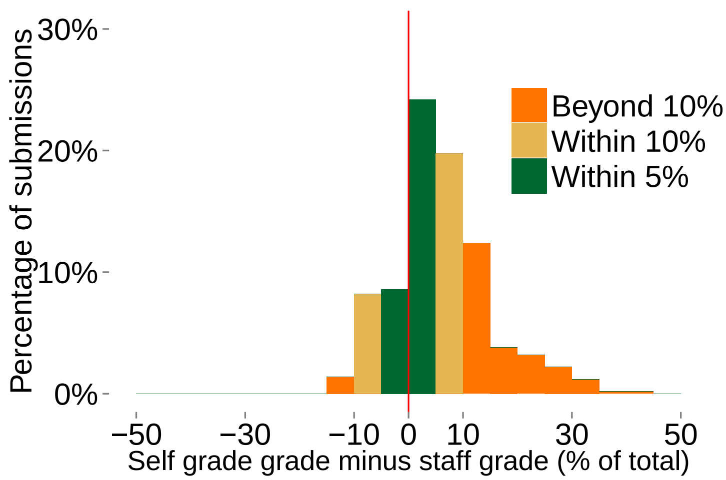


Fig. 6. Agreement of self and staff grades in an in-person class.

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 [Amabile 1982].

3.4. 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 (cs147) 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 6).

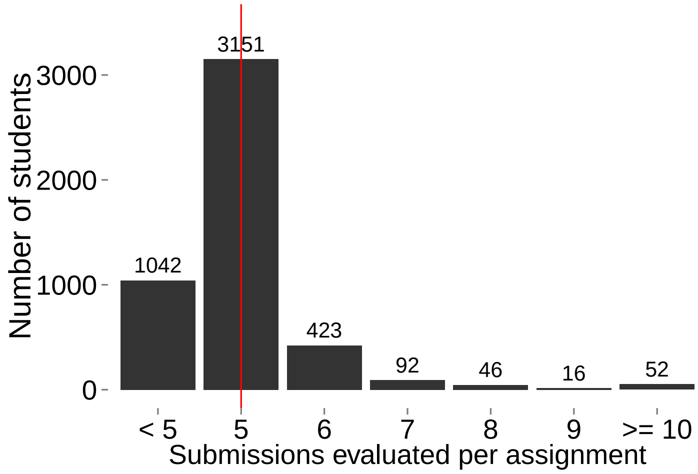


Fig. 7. Average number of submissions assessed per assignment (both iterations). Students were required to assess five, and 20% of students evaluated more than required.

3.5. Results: Student reactions

to see other how other people see how other(s) other's work/other people's	114
points of view point of view	36
compare my work	12
helped me understand	12

(a) **"In what ways was assessing others' work useful?"** Students frequently mentioned being inspired by others work, finding example work to critique, and seeing different points of view.

my own work your own work	175
compare my work I could compare	50
I didn't I did not	31
what I did	19
point of view	15

(b) **"In what ways was assessing your own work useful?"** Students frequently mentioned gaining a new perspective on revisiting their work (after peer assessment), comparing their work to peers', and better identifying their mistakes.

Fig. 8. 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.

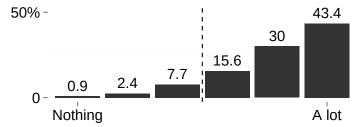
Student reactions to the peer assessment system were generally positive, and 20% of students completed more peer assessments than the class required them to (Figure 7). We infer from this that students found rating their peers valuable or enjoyable, and/or they believed it would help their peers.

42% of students 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 (Figure 8). Overall, students reported

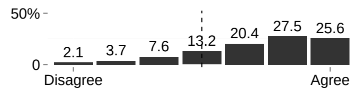
Table III. End course survey results (n=3,550) about student perceptions on peer assessment. Students reported learning from assessing others' work than their own, and putting effort into grading fairly.

Learned from assessing your own work? (Nothing. . . A lot)

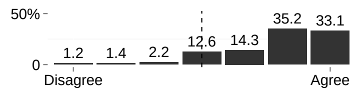
Learned from assessing others' work? (Nothing. . . A lot)



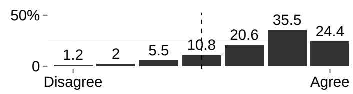
The peer assessment process was easy to understand (Disagree. . . Agree)



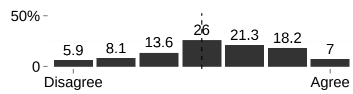
I assessed myself fairly and accurately (Disagree. . . Agree)



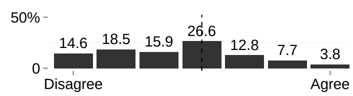
I put sufficient effort into grading peers (Disagree. . . Agree)



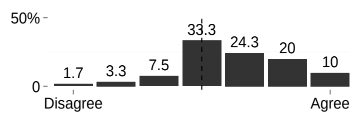
Peers put sufficient effort into grading me (Disagree. . . Agree)



My peer graders did not understand my work (Disagree. . . Agree)



Rubrics helped me understand exactly what assignments required me to do. (Disagree. . . Agree)



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 III). 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 [Ehrlinger et al. 2008]. Designing peer assessment interfaces that emphasize reciprocity and minimize this bias remains future work.

3.6. 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

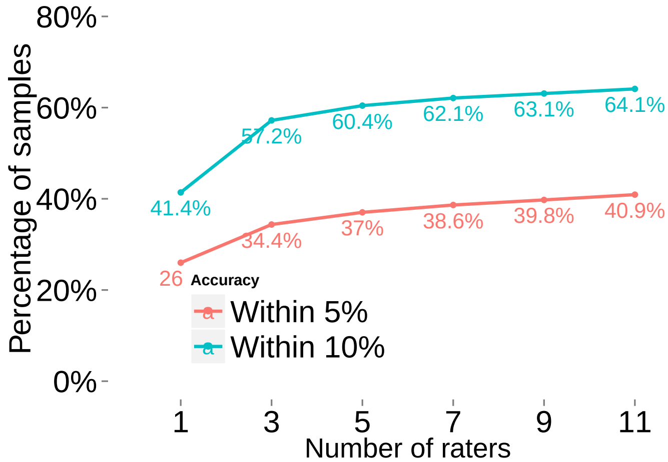


Fig. 9. Increasing the number of raters quickly yields diminishing returns.

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 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 less than 45 students (approx. 6%) were affected by grading errors in each iteration of the class.

3.7. Would more raters help?

Increasing the number of raters per submission helps accuracy, but quickly yields diminishing returns (Figure 9). A large number of students rated staff-graded assignments. 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%.

3.8. 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 10(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 10(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.

3.9. What is the right granularity of grades?

Sections 3.3 and 3.4 shows 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.

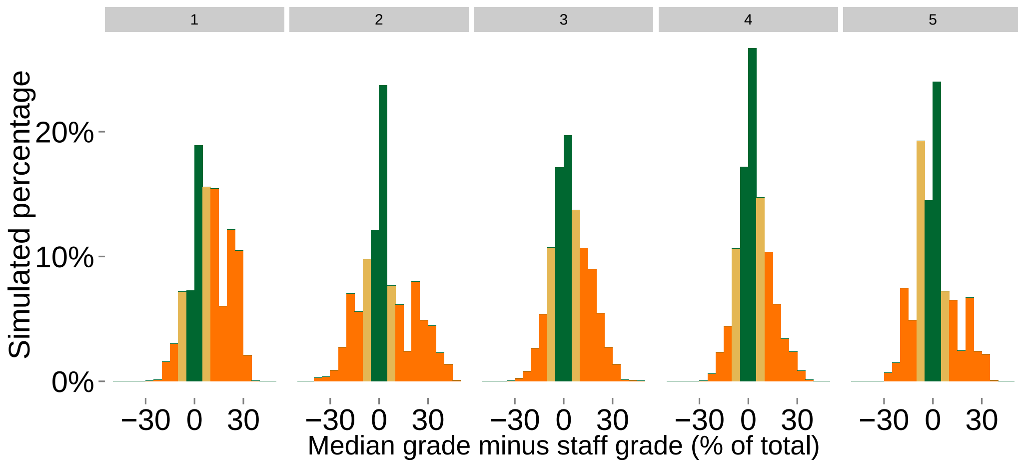
3.10. "Patriotic" grading?

On average, raters grade students from their own country 3.6% higher than those from other countries: $t(27067) = 3.98, p < 0.001$. This effect is consistent when the raters and submitters from the largest student enrollment (United States) are removed, but is smaller (the mean difference drops to 1.98%, $t(12863) = 2.0, p < 0.05$). We remind the reader that grading was double-blind, so raters did not see the names of submitters.

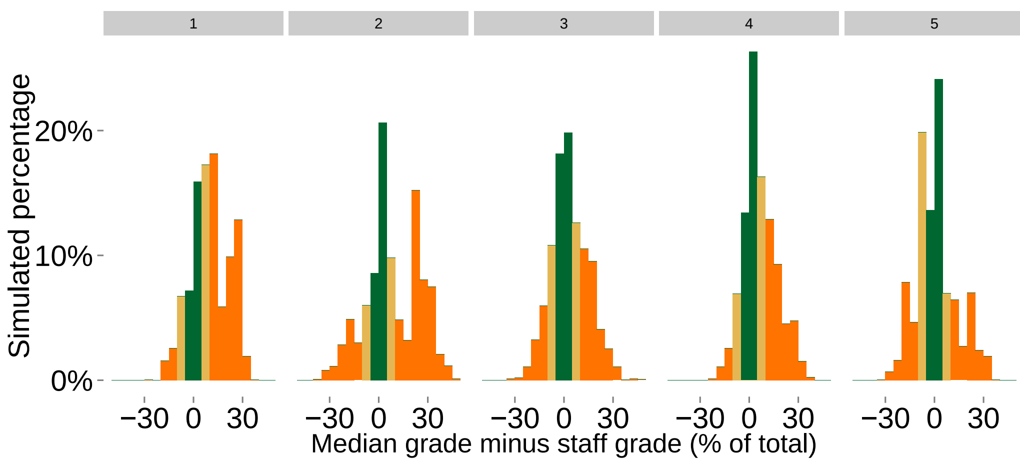
We see four possible explanations for this "patriotism" bias. One is that raters better understood applications designed for their local environment and so rated them more highly. Another is that raters were "voting" for applications that they inferred were from the same country – by the content of the application or the style of the presentation. A third possible explanation is that different cultures consider differing attributes of design, as in Kim and Hinds' work on cross-cultural creativity [Kim and Hinds 2012]. Finally, assessment materials may be understood by students in different countries in subtly different ways. Understanding this effect remains future work.

4. PROVIDING STUDENTS FEEDBACK ON GRADING ACCURACY IMPROVES SUBSEQUENT PERFORMANCE

So far, this paper 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 [Dow et al. 2012], but can it help raters overcome their (possibly unintentional) grading bias? This section describes an experiment that provided students feedback whether they



(a) All raters



(b) Only raters who finished the class

Fig. 10. 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.)

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.

4.1. 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,

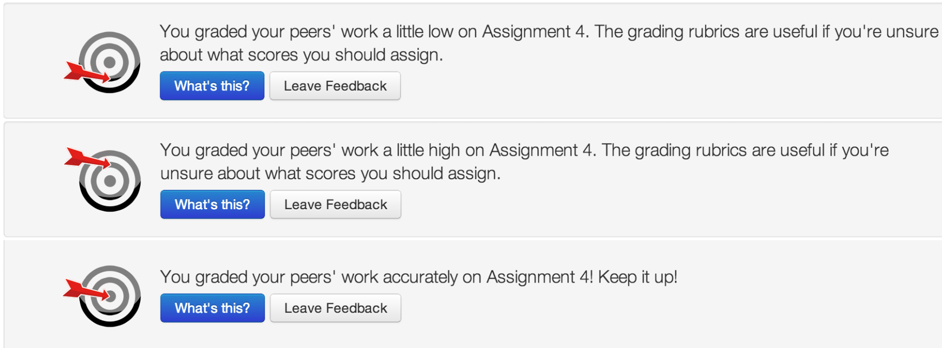


Fig. 11. In the feedback condition, students received feedback about how well they were grading.

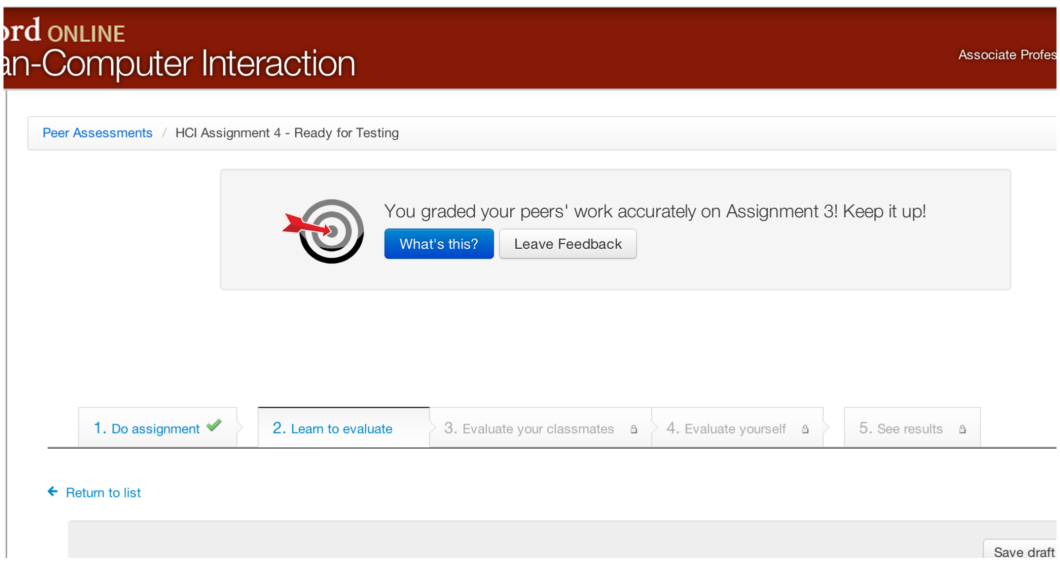


Fig. 12. Students improved grading when provided accuracy feedback.

or just right (Figure 11). 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 12). In the control condition this space was blank.

4.2. 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

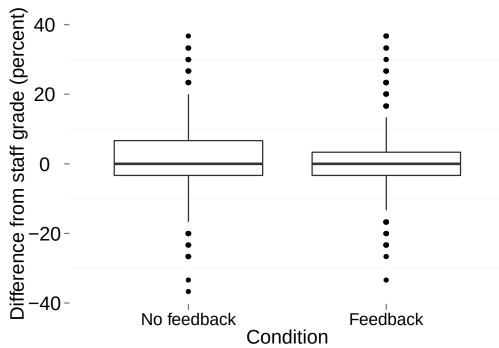


Fig. 13. Feedback on grading accuracy reduced the overall error in assessment and made the range of errors smaller.

effect. 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.

5. PROVIDING PERSONALIZED, QUALITATIVE FEEDBACK ON ASSIGNMENTS

Accurate, actionable feedback helps students improve their work [Nicol and Macfarlane-Dick 2006; Boud 2000]. Actionable feedback is most useful if it is personalized, and targets the student's recent work [Gallien and Oomen-Early 2008].

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 14). Inspired by [Dow et al. 2010], we call the result "fortune-cookie feedback" for its brevity and general applicability. Table IV shows some examples.

Table IV. Example "fortune cookie" feedback

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

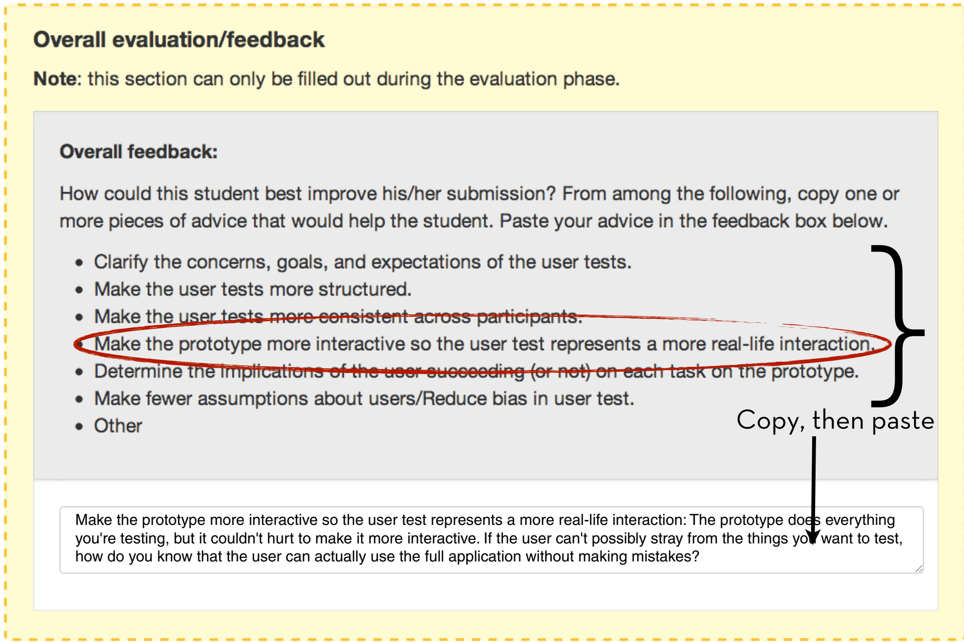


Fig. 14. Students copied snippets of feedback (fortune cookies), pasted them in a textbox and optionally added an explanation.

5.1. 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 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.

5.2. 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 con-

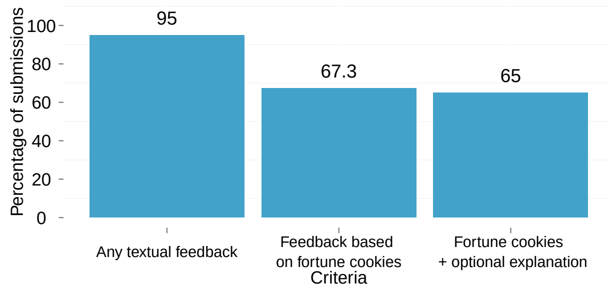


Fig. 15. Most students received at least one piece of textual feedback. Most fortune cookie feedback was personalized

tained feedback suggests that fortune cookies do not encourage more students to leave feedback ($\chi^2 = 0.1, p = 0.75$). Because submissions were assessed by multiple students, 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 15).

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.

5.3. 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 in the current article, 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 [Anderson and Bower 1972; Nielsen 1994]. Second, showing a list of common, assignment-specific problems that the submission could have potentially reduces inhibition, and encourages peers to think critically [Galinsky and Moskowitz 2000]. Third, because fortune cookies sometimes used terminology learned in class, they may have triggered cued-recall of these concepts [Little and Bjork 2012], 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.

6. OVERALL DISCUSSION

6.1. Using data to improve assessment materials

Iterative design often pays big dividends [Nielsen 1993], 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.

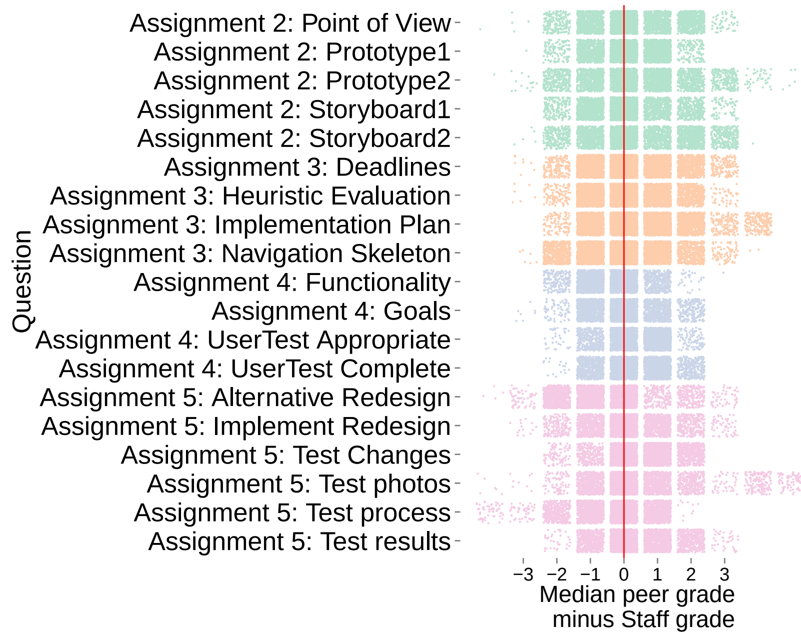


Fig. 16. Comparing variance of rubric items can help teaching staff find areas that may need improvement. For example, this figure shows the variance for four assignments of the HCI course between staff grade and median peer grade. A narrow, dense band indicates higher agreement. For example, Assignment 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 16), 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 around 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. [Andrade 2005]). We hypothesized that using a parallel sentence structure would better help students understand conceptual differences [Markman and Gentner 1993]. 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 17). 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 functional?”).

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).

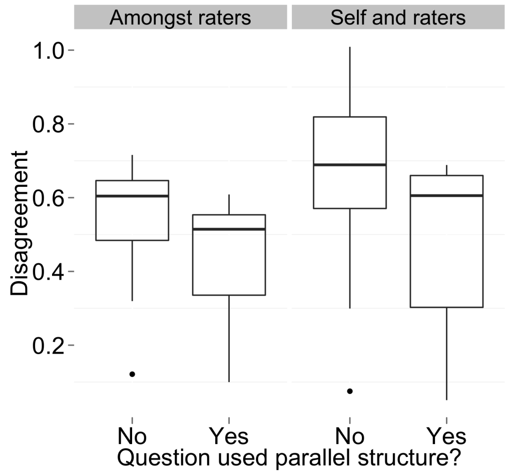


Fig. 17. In iteration 1, questions with parallel structure had lesser disagreement, both amongst peer graders, and between the median grade and the self-assessed grade. We changed all assignments to use parallel structure across rubric items.

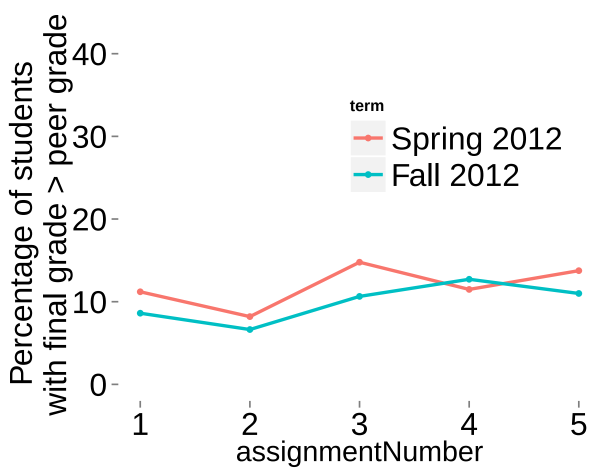


Fig. 18. Students in the second (Fall 2012) iteration of the class reported a self grade $>$ 5% higher than peer grade more frequently, and so got their self grade less frequently.

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.

6.2. Going beyond pass/fail

Peer assessment as described in this paper 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.

6.3. 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 [Mazar et al. 2008]. (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 [Widom 2012]. 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 18). 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 [Boud 1995], 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. [Kurhila 2012; Lewin 2013b]). To combat this, several organizations have proposed solutions like in-person testing facilities (e.g. [Lewin 2012b]), or verified-identity certification [Lewin 2013d]. Others remain focused on teaching for students who want to learn [Widom 2012].

6.4. 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 [Conti et al. 1996]. This may be because novices and experts differ in their tacit understanding of value [Kaufman et al. 2008]. 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.

6.5. 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 [Kuebli et al. 2008]. 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 [Perry 1970]. 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 [Palen 1999].

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.

6.6. 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 [Fox and Patterson 2012]). 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 [Roberts et al. 1995]. 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.

6.7. 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 [Breslow et al. 2013].

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 [Lewin 2013c], we found that campus-paced 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 [Schön 1985; Pendleton-Jullian 2010]. 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 [Klemmer et al. 2006]. In addition, a studio facilitates dialogue between students, instructors and artifacts that helps students collaboratively learn difficult concepts and solve problems [Schön 1985].

The opportunity here is two-fold. First, online learning can be blended with colocated 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 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 [Kizilcec et al. 2013]. 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 tell us is that lots of people are curious. How can we convert this curiosity into meaningful learning opportunities for more students?

7. CONCLUSIONS AND FUTURE WORK

This paper described our experiences with the largest use of peer assessment to date. This paper 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 [Ipeirotis et al. 2010]. Also, the number of raters could be dynamically assigned to be the minimum required for consensus, extending *e.g.* [Guo et al. 2012]. Furthermore, an algorithm could adaptively select particular raters based on estimated quality, focusing high quality work where it's most needed, as in [Dai et al. 2010]. Finally, as with standardized essay grading [Hearst 2000], peers could be used together with automated grading algorithms (such as [Socher et al. 2012; Zaidan and Callison-Burch 2011]). 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 [Corbett et al. 2002]. 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. [Kraut and Resnick 2011]) and resources used at physical universities, such as mentoring and orientation courses [Murtaugh et al. 1999]. 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 [Marlow et al. 2013]. Another is to leverage peers in physical meet-ups to augment instructor teaching [Cadiz et al. 2000]. 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 [Martin 2012]. 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 [Kittur et al. 2013]. 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.

APPENDIX

A.1. 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 19). 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)

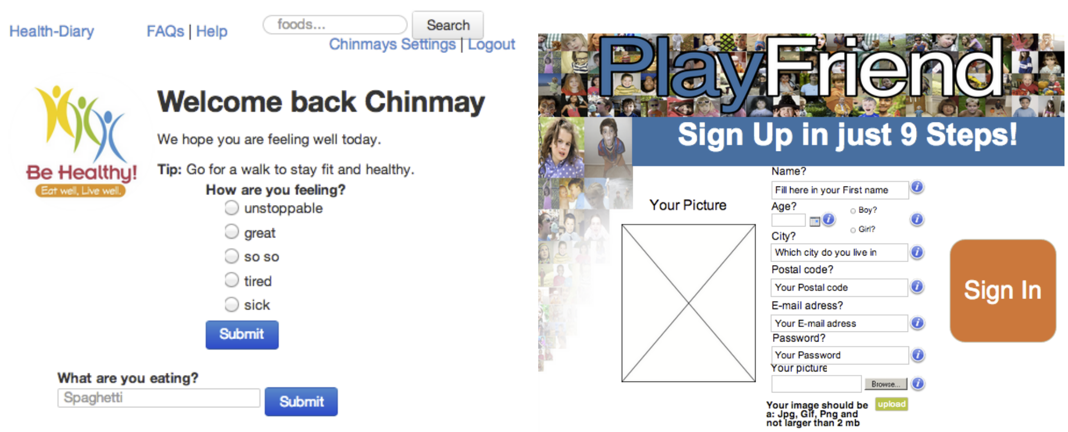
A.2. Grading differences

A.2.1. Where peers graded higher. Figure 20(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."

A.2.2. Where peers graded lower. Figure 20(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",

Fig. 19. Agreement of unaggregated peer grades and staff grades. Agreement is much lower than between *median* peer grades and staff grades.

while some peers rated it "unpolished", and asked the student to "Try to make UI less coloured."



(a) Submission where peers grade higher than staff (b) Submission with staff grade higher than peers

Fig. 20. Student submissions with large differences between staff and peer grades.

B. SAMPLE RUBRIC

Table V shows a rubric for the "Ready for testing" assignment. All other rubrics are available as online supplementary materials.

ACKNOWLEDGMENTS

We thank Coursera for implementing this peer assessment system, and enabling us to use the data. We thank Sébastien Robaszkiewicz, Joy Kim and our Community TAs for helping revise assignments, assess student submissions, and provide forum support; Nisha Masharani for helping collect data and for designing fortune-cookie feedback; and Sébastien Robaszkiewicz and Julie Fortuna for rating fortune cookies. We thank Greg Little for discussions about online peer assessment and Michael Bernstein for comments on drafts. We thank Jane Manning and colleagues at Stanford for supporting this class's development. We thank our editor Marti Hearst and anonymous reviewers for their valuable feedback and suggestions. Human subjects research was reviewed by the Stanford Institutional Review Board through protocol 25001. We thank all of the students in the HCI online class for their ~~foolhardiness~~ enthusiasm in participating in this experimental class.

Author Statement

This submission has no prior publications.

REFERENCES

L. Alben. 1996. Defining the criteria for effective interaction design. *interactions* 3, 3 (1996), 11–15.

T.M. Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology* 43, 2 (1982), 997–1013.

J.R. Anderson and G.H. Bower. 1972. Recognition and retrieval processes in free recall. *Psychological review* 79, 2 (1972), 97–123.

H.G. Andrade. 2005. Teaching with rubrics: The good, the bad, and the ugly. *College Teaching* 53, 1 (2005), 27–31.

R.E. Bennett. 1998. Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice* 17, 4 (1998).

R.E. Bennett, M. Steffen, M.K. Singley, M. Morley, and D. Jacquemin. 1997. Evaluating an Automatically Scorable, Open-Ended Response Type for Measuring Mathematical Reasoning in Computer-Adaptive Tests. *Journal of Educational Measurement* 34, 2 (1997), 162–76.

D. Boud. 1995. *Enhancing learning through self assessment*. Routledge.

D. Boud. 2000. Sustainable assessment: rethinking assessment for the learning society. *Studies in continuing education* 22, 2 (2000), 151–167.

L.B. Breslow, D.E. Pritchard, J. DeBoer, G.S. Stump, A.D. Ho, and D.T. Seaton. 2013. Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment* 8 (2013), 13–25.

B. Buxton. 2007. *Sketching user experiences: getting the design right and the right design*. Morgan Kaufmann.

JJ Cadiz, A. Balachandran, E. Sanocki, A. Gupta, J. Grudin, and Gavin Jancke. 2000. Distance learning through distributed collaborative video viewing. In *ACM conference on Computer supported cooperative work*. ACM, 135–144.

P.A. Carlson and F.C. Berry. 2003. Calibrated Peer Review and assessing learning outcomes. In *Frontiers in Education Conference*, Vol. 2. STIPES.

S. Carter, J. Mankoff, S.R. Klemmer, and T. Matthews. 2008. Exiting the cleanroom: On ecological validity and ubiquitous computing. *Human–Computer Interaction* 23, 1 (2008), 47–99.

K. Cennamo, S. A Douglas, M. Vernon, C. Brandt, B. Scott, Y. Reimer, and M. McGrath. 2011. Promoting creativity in the computer science design studio. In *Proceedings of the 42nd ACM technical symposium on Computer science education*. ACM, 649–654.

C. Cheshire and J. Antin. 2008. The social psychological effects of feedback on the production of Internet information pools. *Journal of Computer-Mediated Communication* 13, 3 (2008), 705–727.

E.H. Chi. 2009. 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*. Springer-Verlag, 597–605.

D. Chinn. 2005. Peer assessment in the algorithms course. *ACM SIGCSE Bulletin* 37, 3 (2005), 69–73.

R. Conti, H. Coon, and T.M. Amabile. 1996. Evidence to support the componential model of creativity: Secondary analyses of three studies. *Creativity Research Journal* 9, 4 (1996), 385–389.

A. T Corbett, K.R. Koedinaer, and W. Haaley. 2002. Cognitive Tutors: From the Research Classroom to All Classrooms. *Technology enhanced learning: opportunities for change* (2002), 235.

P. Dai, Mausam D., and D.S. Weld. 2010. Decision-theoretic control of crowd-sourced workflows. In *In the 24th AAAI Conference on Artificial Intelligence (AAAI10)*. Citeseer.

D. P Dannels and K.N. Martin. 2008. Critiquing Critiques A Genre Analysis of Feedback Across Novice to Expert Design Studios. *Journal of Business and Technical Communication* 22, 2 (2008), 135–159.

B. De La Harpe, J.F. Peterson, N. Frankham, R. Zehner, D. Neale, E. Musgrave, and R. McDermott. 2009. Assessment focus in studio: What is most prominent in architecture, art and design? *International Journal of Art & Design Education* 28, 1 (2009), 37–51.

S.P. Dow, A. Glassco, J. Kass, M. Schwarz, D.L. Schwartz, and S.R. Klemmer. 2010. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction (TOCHI)* 17, 4 (2010), 18.

S. Dow, A. Kulkarni, S. Klemmer, and B. Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1013–1022.

A. Drexler, R. Chafee, and others. 1977. *The Architecture of the Ecole des Beaux-Arts*. distributed by MIT Press.

B. Efron and R. Tibshirani. 1993. *An introduction to the bootstrap*. Vol. 57. Chapman & Hall/CRC.

J. Ehrlinger, K. Johnson, M. Banner, D. Dunning, and J. Kruger. 2008. Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes* 105, 1 (2008), 98–121.

N. Falchikov and J. Goldfinch. 2000. Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research* 70, 3 (2000), 287–322.

D. Fallman. 2003. Design-oriented human-computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 225–232.

E.B. Feldman. 1994. *Practical art criticism*. Prentice Hall New York.

J. Forlizzi and K. Battarbee. 2004. Understanding experience in interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*. ACM, 261–268.

A. Fox and D. Patterson. 2012. Crossing the software education chasm. *Commun. ACM* 55, 5 (2012), 44–49.

A. D Galinsky and G. B Moskowitz. 2000. Counterfactuals as behavioral primes: Priming the simulation heuristic and consideration of alternatives. *Journal of Experimental Social Psychology* 36, 4 (2000), 384–409.

T. Gallien and J. Oomen-Early. 2008. 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* 7, 3 (2008), 463–476.

R.D. Gerdeman, A.A. Russell, and K.J. Worden. 2007. Web-Based Student Writing and Reviewing in a Large Biology Lecture Course. *Journal of College Science Teaching* 36, 5 (2007), 46–52.

S. Greenberg. 2009. Embedding a design studio course in a conventional computer science program. In *Creativity and HCI: From Experience to Design in Education*. Springer, 23–41.

S. Guo, A. Parameswaran, and H. Garcia-Molina. 2012. So who won?: dynamic max discovery with the crowd. In *Proceedings of the 2012 international conference on Management of Data*. ACM, 385–396.

M. A Hearst. 2000. The debate on automated essay grading. *Intelligent Systems and their Applications, IEEE* 15, 5 (2000), 22–37.

S. Hsi and A. M Agogino. 1995. Scaffolding knowledge integration through designing multimedia case studies of engineering design. In *Frontiers in Education Conference, 1995. Proceedings., 1995*, Vol. 2. IEEE, 4d1–1.

S.W. Huang and W.T. Fu. 2013. Enhancing Reliability Using Peer Consistency Evaluation in Human Computation. In *Proceedings of ACM: Computer supported collaborative work*. ACM.

P.G. Ipeirotis, F. Provost, and J. Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 64–67.

J.C. Kaufman, J. Baer, J.C. Cole, and J.D. Sexton. 2008. A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal* 20, 2 (2008), 171–178.

F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, and others. 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology* 18, 10 (2011), 1175–1177.

H. Kim and P. Hinds. 2012. 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*. ACM.

A. Kittur, J. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton. 2013. The future of crowd work. In *ACM Conference on Computer Supported Coooperative Work (CSCW 2013)*.

R.F. Kizilcec, C. Piech, and E. Schneider. 2013. Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. (2013), 170–179.

S.R. Klemmer, B. Hartmann, and L. Takayama. 2006. How bodies matter: five themes for interaction design. In *Proceedings of the 6th conference on Designing Interactive systems*. ACM, 140–149.

R.E. Kraut and P. Resnick. 2011. *Evidence-based social design: Mining the social sciences to build online communities*. MIT Press.

J.E. Kuebli, R.D. Harvey, and J.H. Korn. 2008. Critical Thinking in Critical Courses: Principles and Applications. *Teaching Critical Thinking in Psychology: A Handbook of Best Practices* (2008), 137.

J. Kurhila. 2012. Human-Computer Interaction by Coursera opened for credit for the students of the Department. http://www.cs.helsinki.fi/en/uutiset/72025, (2012).

B. Lawson. 2006. *How designers think: the design process demystified*. Architectual Press.

T. Lewin. 2012a. *Education Site Expands Slate of Universities and Courses*. The New York Times.

T. Lewin. 2012b. One Course, 150,000 Students. (7 2012).

T. Lewin. 2013a. *College of Future Could Be Come One, Come All*. January 2013, The New York Times.

T. Lewin. 2013b. *Five Online Courses Are Eligible for College Credit*. February 2013, The New York Times.

T. Lewin. 2013c. *Students Rush to Web Classes, but Profits May Be Much Later*. January 2013, The New York Times.

T. Lewin. 2013d. *Universities Abroad Join Partnerships on the Web*. February 2013, The New York Times.

J.L. Little and E.L. Bjork. 2012. Pretesting with Multiple-choice Questions Facilitates Learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

A.B. Markman and D. Gentner. 1993. Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language* 32 (1993), 517–517.

J. Marlow, L. Dabbish, and J. Herbsleb. 2013. Impression formation in online peer production: activity traces and personal profiles in github. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 117–128.

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

N. Mazar, O. Amir, and D. Ariely. 2008. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research* 45, 6 (2008), 633–644.

P.A. Murtaugh, L.D. Burns, and J. Schuster. 1999. Predicting the retention of university students. *Research in Higher Education* 40, 3 (1999), 355–371.

D.J. Nicol and D. Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education* 31, 2 (2006), 199–218.

J. Nielsen. 1993. Iterative user-interface design. *Computer* 26, 11 (1993), 32–41.

J. Nielsen. 1994. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 152–158.

L. Palen. 1999. 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*. ACM, 17–24.

A. Pendleton-Jullian. 2010. *Four (+1) Studios*. CreateSpace Independent Publishing.

W.G. Perry. 1970. Forms of intellectual development in the college years. *New York: Holt* (1970).

P.R. Pintrich. 1995. Understanding self-regulated learning. *New directions for teaching and learning* 1995, 63 (1995), 3–12.

P. Pintrich and A. Zusho. 2007. Student motivation and self-regulated learning in the college classroom. *The scholarship of teaching and learning in higher education: An evidence-based perspective* (2007), 731–810.

Y.J. Reimer and S.A. Douglas. 2003. Teaching HCI design with the studio approach. *Computer Science Education* 13, 3 (2003), 191–205.

E. Roberts, J. Lilly, and B. Rollins. 1995. Using undergraduates as teaching assistants in introductory programming courses: An update on the Stanford experience. *ACM SIGCSE Bulletin* 27, 1 (1995), 48–52.

D. Schön. 1985. The Design Studio: An exploration of its traditions and potential. *London: Royal Institute of British Architects* (1985).

A. Snodgrass and R. Coyne. 2006. *Interpretation in architecture: Design as a way of thinking*. Routledge.

R. Socher, B. Huval, C.D. Manning, and A.Y. Ng. 2012. Semantic Compositionality Through Recursive Matrix-Vector Spaces. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

C.A. Stanley and M.E. Porter. 2002. *Engaging Large Classes: Strategies and Techniques for College Faculty.* ERIC.

J. Surowiecki. 2005. *The wisdom of crowds*. Anchor.

M. Szpir. 2002. Clickworkers on Mars. *American Scientist* 90, 3 (2002).

D. Tinapple, L. Olson, and John Sadauskas. 2013. CritViz: Web-Based Software Supporting Peer Critique in Large Creative Classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology* 15, 1 (2013), 29.

M. Tohidi, W. Buxton, R. Baecker, and A. Sellen. 2006. Getting the right design and the design right. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 1243–1252.

J.E. Tomayko. 1991. Teaching software development in a studio environment. *ACM SIGCSE Bulletin* 23, 1 (1991), 300–303.

K. Topping. 1998. Peer assessment between students in colleges and universities. *Review of Educational Research* 68, 3 (1998), 249–276.

B. Uluoglu. 2000. Design knowledge communicated in studio critiques. *Design Studies* 21, 1 (2000), 33–58.

A. Venables and R. Summit. 2003. Enhancing scientific essay writing using peer assessment. *Innovations in Education and Teaching International* 40, 3 (2003), 281–290.

J. Widom. 2012. *From 100 Students to 100,000*. ACM SIGMOD Blog (http://wp.sigmod.org/?p=165).

T. Winograd. 1990. What can we teach about human-computer interaction?(plenary address). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 443–448.

O.F. Zaidan and C. Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. 1220–1229.

B.J. Zimmerman and D.H. Schunk. 2001. Reflections on theories of self-regulated learning and academic achievement. *Self-regulated learning and academic achievement: Theoretical perspectives* 2 (2001), 289–307.

Table V. Rubric for “Ready for Testing” assignment. Students have created a paper prototype of their application in the previous assignment. Note some items have objective criteria (Did the student meet her goals?), others require subjective interpretation (Is this evaluation plan appropriate?)

Category	Unsatisfactory	Bare minimum	Satisfactory effort & performance	Above & Beyond
List of Changes	0: No changes or completely irrelevant changes.	1: The student only identified a few changes from the heuristic evaluation feedback and a large amount of feedback is ignored in the new prototype; the new prototype has some HE violations.	3: Many of the simpler suggested changes were made, but some of the more complex or difficult issues were not addressed; the new prototype does not have any obvious HE violations.	5: The user made several insightful and specific changes based on the heuristic evaluation feedback. It is hard to find any HE violations at all in the new prototype.
Interactive Prototype	0: No prototype or irrelevant prototype.	1: The prototype is not interactive, lacks many features, and has many bugs; the design does not work with the goal. OR, the student submitted a prototype URL, but the prototype wasn't viewable.	3: The prototype is mostly interactive, with only a few features missing and only one or two bugs; the design accomplishes the minimum requirements of the goal..	5: The prototype is completely interactive, reflects the feel of the final prototype, and is ready for user testing; the design accomplishes the entire goal.
User Evaluation Plan: Completeness	0: No plan or irrelevant plan.	1: User testing evaluation plan exists, but is minimal, unclear, and is not well thought out.	3: The evaluation plan is mostly complete, but does not cover all questions about testing thoroughly (what is tested, what you want to learn, when, where, participants).	5: The evaluation plan is complete, answers all questions specifically, and shows a clear process for user testing.
User Evaluation Plan: Appropriateness	0: No plan or irrelevant plan.	1: The student's evaluation plan does not choose to evaluate aspects of the design related to the design goals.	3: The evaluation plan is designed to produce some useful data, but is not justified by the student (e.g. why are you doing what you are doing?– why 6 participants? Why in a school? etc).	5: The evaluation plan is very clearly motivated or innovative in a way that will ensure rich and interesting data to address the design goals.
Development Goals	0: No goals met that were laid out on the development plan.	1: The student met a few of the goals laid out in the development plan.	2: The student met most, but not all, of the goals laid out in the development plan.	3: The student met all of the goals found in the development.