Long-Term Peer Reviewing Effort is Anti-Reciprocal

Yasmine Kotturi[1], **Andrew Du**[2], **Scott Klemmer**[2], **Chinmay Kulkarni**[1]
[1]HCI Institute, Carnegie Mellon, [2]Design Lab, UC San Diego
{ykotturi, chinmayk}@cs.cmu.edu, {aadu, srk}@ucsd.edu

ABSTRACT
Many studies demonstrate that peer reviewing provides pedagogical benefits such as inspiration and developing expert vision, and changes classroom culture by encouraging reciprocity. However, much large-scale research in peer assessment has focused on MOOCs, where students have short tenures, and is unable to describe how reciprocity-oriented classroom cultures evolve over time. This short paper presents the first long-term analysis of peer reviewing with 304 students, conducted in three large physical classes in a year-long undergraduate series. Surprisingly, this analysis reveals that when students receive better reviews on their work, they write worse reviews in the future. This suggests that while students believe in the reciprocal nature of peer review, they act anti-reciprocally. Therefore, battling the emergent norm of anti-reciprocity is crucial both for system designers and practitioners who use peer assessment.

Author Keywords
peer assessment, peer review, reciprocity

INTRODUCTION
Peer assessment can be pedagogically powerful: it exposes students to ideas, supports reflection, enhances critical thinking, improves course performance, and decreases attrition [1, 4, 14, 16, 17]. Peer review at scale can also catalyze fast, formative feedback on in-progress open ended work at massive scale [12]. Such feedback and iteration are important for mastery learning [6].

However, why do students put in substantial effort in helping their peers, rather than the minimum effort required? Thus far, reciprocity, i.e. our desire to help those who help us [7], has been cited as a likely reason [2]. Does this desire diminish over time, or does seeing the others provide help build a stronger norm, strengthening reciprocity over the long-term? We believe the answer to this question is of fundamental importance to the long-term sustainability of peer review as an academic practice. Unfortunately, current large-scale research on peer review is unable to answer this fundamental question, because it has largely been conducted in large scale settings where students have short tenures (e.g. [8, 10, 11, 12, 13, 18]). These short tenures are insufficient to study how peer assessment evolves over time.

L@S 2017, April 20-21, 2017, Cambridge, MA, USA
 ISBN 978-1-4503-4450-0/17/04…$15.00
DOI: http://dx.doi.org/10.1145/3051457.3054004

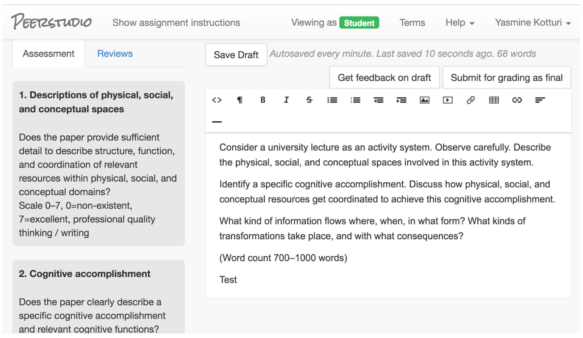

Figure 1. PeerStudio enables students to get fast peer feedback and iterate on their open-ended work.

We introduce the first long-term analysis, one academic year (2015-2016), of peer reviewing behavior in three large, physical classes. In total, we analyze data from 304 active students as they progress through all three classes which comprise an undergraduate year long series; each class used our peer review platform, PeerStudio (Figure 1) [12]. These analyses suggest an emergent norm of anti-reciprocity, where over time students perform reviews of declining quality.

Course and Assignment Format
We analyze data from 304 students who participated in a three-course series in UC San Diego's Cognitive Science undergraduate program, taught by the same instructor. Students learn about topics such as distributed cognition and cognitive ethnography. Students wrote essays (between 500-1000 words) for assignments, such as reviews of research articles. Rubrics evaluating these assignments comprised a number of Likert scales. Overall, our dataset comprises 10,845 reviews generated on 4,131 submissions.

Hypotheses
Our conversations with the course staff teaching this series revealed an increasing frustration with peer assessment over time. Often, course staff would ask for best practices to combat what they saw as an increasing lack of student interest over time. Based on these observations, we hypothesized that:

(H1) Review quality decreases over time.

Since peer reviewing is a reciprocal act, we believed that students' review quality (H1) was likely the result of visible norms: when students saw that poor reviewing went sans

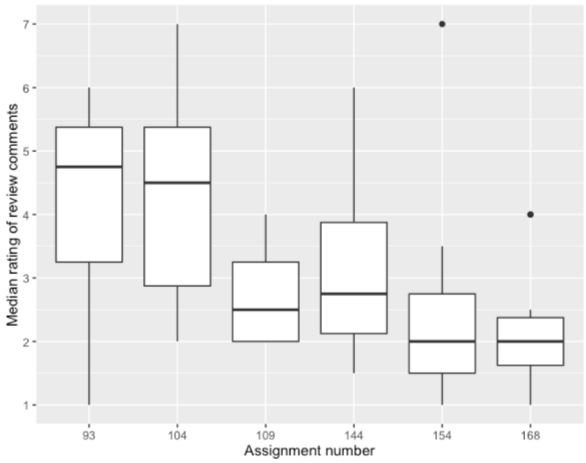

Figure 2. Review quality decreases as students progress through year long course series. Coders rated quality of 60 comments on 7-pt scale [3]. Assignments 93 and 104 are the first and last assignments in the first course, respectively. Assignments 109 and 144 are first and last in the second course, and so on. Assignments were about 6 weeks apart.

repercussion, they were less likely to put in effort themselves [9]. This leads to our second hypothesis:

(H2) Reviewers act on reciprocity: when students receive a good review, they are more likely to generate a good review on the next assignment.

Measures

We analyzed review comment length, quality of reviews, and their potential interaction. To analyze the quality of a review, we extracted a subset of 60 reviews: 10 reviews from the first and last assignments in each course. Four blinded-to-assignment coders–HCI graduate students–(1F/3M, ages 22-29), rated this subset of reviews on a continuous scale adapted from prior work [3]; from one to seven with the following coding schema: (1) Irrelevant or no discussion: review offers no actionable feedback (4) Merely stated edits: review suggests edits and changes, gives some justification or reasoning (7) Actionable and justified: Review articulates clear actionable edits with sound and succinct justification. Because our raters had fair agreement on this scale, we use the median rating as the aggregate.

RESULTS

We first analyzed review comment length from all 10,845 reviews and we saw that, generally, length decreased over time (Figure 3). We then examined the quality of reviews, as rated by the four coders. For the 60 rated reviews, we see that quality decreased over time (Figure 2). We therefore investigated the interaction between review length and quality.

Review quality increases with comment length

Across the 60 reviews that our coders rated, we found that the word length of the review was highly correlated with the median rating of quality (Spearman $\rho = 0.88$). Because the review length is much easier to compute, we use it as a proxy for comment quality for further large-scale analyses.

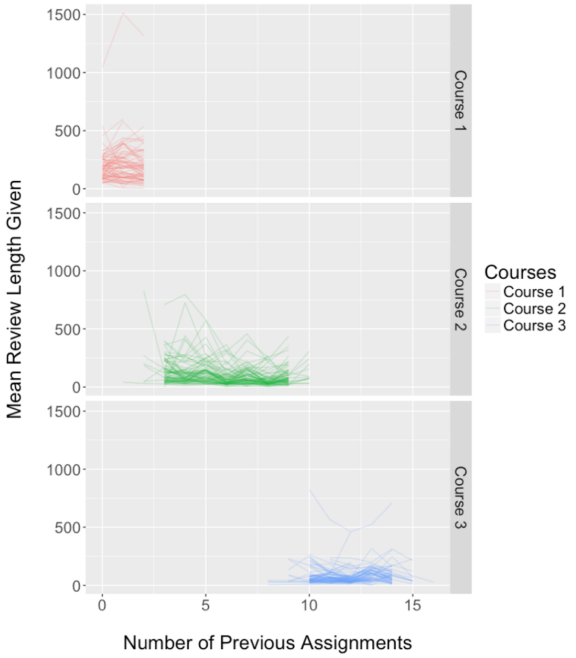

Figure 3. Data from 10,845 reviews: as number of assignments completed increases review length decreases.

Coefficients	β	F	p-value
Intercept	86.14165	17.74	<0.001
Previous Mean Review Length Given	0.61138	42.58	<0.001
Second Course	-54.91843	-12.08	<0.001
Third Course	-57.35288	-11.14	<0.001

Table 1. Students' previous reviewing quality significantly affects their current reviewing quality.

R1: Review quality decreases over time

To understand how review quality changes over time, we built a linear regression model that used the word length of a student's previous review to predict the length of their current review. The course name was included as a fixed effect covariate (since courses were chosen to be part of a series). This model has strong fit ($R^2 = 0.47$), and students' previous reviewing quality significantly affects their current reviewing quality ($F(1,2505) = 48.04, p < 0.001$) (See Table 1). On average, students write reviews that are only 66.9% as long as their reviews for the previous assignment. Adding the course as a covariate to the model significantly improves model fit ($R^2 = 0.51$), suggesting that review quality has significant variation within courses in the same series. On average, students write approximately 50 fewer words for each new course in the series. We then examined patterns of review behavior, based on the types of review students received. We see that within course variation in length of review is much lower than variation across courses (Figure 4).

R2: Changes in review quality are anti-reciprocal

To understand more specifically how the quality of reviews written by a student are influenced by the reviews they re-

Coefficients	β	F	p-value
Intercept	115.05531	18.88	<0.001
Previous mean review length given	0.63564	43.72	<0.001
Previous mean review length received	-0.16359	-7.70	<0.001
Second Course	-67.86219	-14.14	<0.001
Third Course	-74.16512	-13.39	<0.001

Table 2. When students see longer reviews on their previous work, they write shorter reviews

ceive, we added the average word length of reviews received on their previous submission as a covariate. Doing so slightly improves model fit ($R^2 = 0.52$), suggesting others reviewing significantly affects reviewing behavior. However, the direction of this influence is surprising. When students see longer reviews on their previous work, they write shorter reviews (See Table 2). On average, for every additional word in reviews students receive, they write 0.16 fewer words in their own reviewing: $(t(2502) = -7.70, p < 0.001)$.

Adding an interaction variable (between previous comment length written, and previous comment length received) shows no significant interaction $(t(2501) = 0.12, p > 0.8)$. This suggests that even the most motivated students (who initially write reviews of high quality) nonetheless write shorter (worse) reviews when they see longer (better) reviews on their own work.

Implications for Design

Our finding also has implication for the design of peer assessment systems: we suggest that relying exclusively on reciprocal social nature may be an insufficient design lever. Instead, other motivational methods, such as increasing social translucence [5], visible monitoring of students' progress, not only on assignments, but also on review quality, and more tightly integrating the platform with physical course activities [8]; for instance, highlighting positive review behavior by decomposing examples of good reviews during class meetings, or, if online, in class announcements.

DISCUSSION AND FUTURE WORK

Current peer review systems rely on student reciprocity. Our findings (admittedly based on one course-series at one university) offer preliminary evidence that peer review systems may lead to anti-reciprocal behavior in the long-term. While particular design features of the system may be responsible, we speculate our finding may represent more than a "bug" in the design of current peer assessment systems. Instead, we speculate three fundamental causal pathways, which future work could investigate. First, students who see well-formed reviews perhaps see the standard as unachievably high, and therefore do not put in further effort reviewing. We see similar results when students are shown extremely high-quality peer work as inspiration [15]. Second, seeing high quality reviews may encourage a diffusion of responsibility. In essence, once students see someone else is working hard on reviewing, they may believe they need to carry a smaller burden [19]. Third,

Figure 4. Students write shorter reviews when they see longer reviews on their own work

over time students who do not receive recognition for their high efforts may gradually become less driven to continue to generate high quality reviews. This work was conducted under IRB protocol #140267XX and was partially funded via NSF grant #IIS-0745320.

REFERENCES

1. David Boud, Ruth Cohen, and Jane Sampson. 2014. *Peer learning in higher education: Learning from and with each other*. Routledge.
2. Kwangsu Cho and Christian D Schunn. 2007. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education* 48, 3 (2007), 409–426.
3. Derrick Coetzee, Seongtaek Lim, Armando Fox, Bjorn Hartmann, and Marti A Hearst. 2015. Structuring interactions for large-scale synchronous peer learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1139–1152.
4. Catherine H Crouch and Eric Mazur. 2001. Peer instruction: Ten years of experience and results. *American journal of physics* 69, 9 (2001), 970–977.
5. Thomas Erickson and Wendy A Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM transactions on computer-human interaction (TOCHI)* 7, 1 (2000), 59–83.
6. K Anders Ericsson and Paul Ward. 2007. Capturing the naturally occurring superior performance of experts in the laboratory toward a science of expert and exceptional

performance. *Current Directions in Psychological Science* 16, 6 (2007), 346–350.

7. Alvin W Gouldner. 1960. The norm of reciprocity: A preliminary statement. *American sociological review* (1960), 161–178.
8. Yasmine Kotturi, Chinmay E Kulkarni, Michael S Bernstein, and Scott Klemmer. 2015. Structure and messaging techniques for online peer learning systems that increase stickiness. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. ACM, 31–38.
9. Robert E Kraut, Paul Resnick, Sara Kiesler, Moira Burke, Yan Chen, Niki Kittur, Joseph Konstan, Yuqing Ren, and John Riedl. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.
10. Chinmay Kulkarni, Julia Cambre, Yasmine Kotturi, Michael S Bernstein, and Scott R Klemmer. 2015b. Talkabout: Making distance matter with small groups in massive classes. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1116–1128.
11. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R Klemmer. 2015c. Peer and self assessment in massive online classes. In *Design thinking research*. Springer, 131–168.
12. Chinmay E Kulkarni, Michael S Bernstein, and Scott R Klemmer. 2015a. PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. ACM, 75–84.
13. Tricia Ngoon, Alexander Gamero-Garrido, and Scott Klemmer. 2016. Supporting Peer Instruction with Evidence-Based Online Instructional Templates. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 301–304.
14. Leo Porter, Cynthia Bailey Lee, and Beth Simon. 2013. Halving fail rates using peer instruction: a study of four computer science courses. In *Proceeding of the 44th ACM technical symposium on Computer science education*. ACM, 177–182.
15. Todd Rogers and Avi Feller. 2016. Discouraged by peer excellence: Exposure to exemplary peer performance causes quitting. *Psychological science* 27, 3 (2016), 365–374.
16. Donald Schön. 1987. Educating the reflective practitioner. (1987).
17. Michelle K Smith, William B Wood, Wendy K Adams, Carl Wieman, Jennifer K Knight, Nancy Guild, and Tin Tin Su. 2009. Why peer discussion improves student performance on in-class concept questions. *Science* 323, 5910 (2009), 122–124.
18. Thomas Staubitz, Dominic Petrick, Matthias Bauer, Jan Renz, and Christoph Meinel. 2016. Improving the Peer Assessment Experience on MOOC Platforms. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. ACM, 389–398.
19. Milena Tsvetkova and Michael W Macy. 2014. The social contagion of generosity. *PloS one* 9, 2 (2014), e87275.