Public Peer Review Motivates Higher Quality Feedback

Xu Wang, Carnegie Mellon University, xuwang@andrew.cmu.edu
Yali Chen, Cengage, Carnegie Mellon University, yali.chen@cengage.com
Amanda Godley, University of Pittsburgh, agodley@pitt.edu
Carolyn Rosé, Carnegie Mellon University, cprose@cs.cmu.edu

Abstract: The role of feedback in learning has been well researched, but in practice high quality feedback may be scarce, for example when the source of feedback is from peer learners. Nevertheless, peer feedback may be the main source of formative feedback available in some settings, such as in Massive Open Online Courses (MOOCs). A key part of the problem may be that students do not have sufficient incentive to offer their best feedback in settings where supervision is minimal. In this paper, we investigate whether students provide feedback of higher quality when it is done in a public setting rather than in a private setting. We report on an experimental study with 65 participants randomly assigned to a public feedback and a private feedback condition. We report the effect of the manipulation in terms of the quality of feedback offered as measured by a validated coding scheme, the subjective rating of the feedback, the effect on propensity to revise and success at increasing the quality of the writing. Limitations of the study and implications for practice are discussed.

Introduction

Much research in the Learning Sciences has investigated the properties of effective feedback and the way it plays into the learning process. For example, Graham et al. (2015) found that feedback significantly enhances students' performance on writing; The Knowledge-Learning-Instruction framework (Koedinger et al., 2012) considered timely feedback to be a critical instructional activity across disciplines, from English language learning to science and math; The ICAP framework (Chi & Wylie, 2014) classified peer feedback as an interactive activity, which is a category of activity designated as particularly conducive to learning. Despite the importance of feedback for learning, lack of available high-quality feedback is sometimes the norm, for example in at-scale online learning environments such as Massive Open Online Courses (MOOCs) (Kulkarni et al., 2015; Hicks et al., 2016; Joyner et al., 2016). Research related to the improvement of peer review processes has not always leveraged best practices in feedback research from the Learning Sciences or measured success in terms of qualities of feedback known to be important for learning. On the one hand, questions have been raised regarding a lack of student motivation in writing feedback, resulting in feedback of inconsistent quality (Suen, 2014). However, it is an open question of whether increasing motivation alone would increase the quality of feedback provided by students when feedback is evaluated using best practice rubrics grounded in the Learning Sciences. In this paper, we address the practical question of how to increase availability of high quality feedback by testing the effect of a public peer review paradigm in at-scale learning environment. We measure the impact of the manipulation on the quality of feedback offered, the subjective rating of the feedback, the effect on propensity to revise and success at increasing the quality of the writing. In this way, we take a first step towards bridging work in peer feedback from practice-oriented communities with basic research on feedback and writing from the Learning Sciences. The long-term goal is to leverage students in online learning settings as a resource for one another. Massive online learning settings pose challenges, but with proper scaffolding the scale may eventually serve as an advantage rather than an impediment.

Peer review has become increasingly popular in MOOCs since instructors are not able to grade and provide feedback for the large number of assignments turned in by students. However, in a near anonymous online learning environment, there are few consequences for high or low-quality feedback and the social distance makes it less likely that students are willing to invest substantial effort in offering feedback. Similarly, prior work has found that when forms of discussion identified as valuable for learning in offline settings occur in MOOCs, they are associated with learning as expected, but rarely occur without support (Wang et al., 2016). On the other hand, Wen et al. (2018) found that feedback exchanges in a public discussion forum prior to separation into collaborative groups is more valuable as preparation for collaboration than feedback delivered privately within the collaborative groups once they are formed. This result in particular suggests that in a public forum, either students are motivated to give better feedback or that students benefit from exposure to a wider variety of work and feedback in a more public context. Tinapple et al. (2013) also points to the importance of accountability when providing peer feedback, which may also be a key factor in a public environment.

In traditional peer review scenarios, which are typically more private, students' exposure to classmates' work and feedback is limited to those from the few classmates they have been matched with. Similar to Wen et al. (2018), we explore a new peer review paradigm, where students post their work to a public discussion forum, and then write feedback to a small number of other students in the forum, thus getting full exposure to all assignments and reviews in the class. We investigate whether the increased accountability of the public environment results in higher quality feedback being exchanged when scaffolding is introduced to focus the effort on forms of feedback known to be beneficial. Prior work deconstructing types of feedback (such as summary, praise, problem, solution, etc.) has found specifically that summarizing the peer's ideas, identifying problems and offering solutions in feedback on writing is associated with better revision, and thus considered to be preferable characteristics of feedback (Nelson & Schunn, 2009). In our study, we specifically point students to these aspects of feedback in our instruction and investigate whether students in a public setting are more conscientious of the instruction provided, and go on to provide more high-quality feedback. This study contributes new knowledge about which aspects of feedback (namely, summary, praise, problem and solution) can be manipulated through raising the level of perceived supervision as a result of situating the feedback activity in an apparently public setting. The study also reveals the extent to which raising the prevalence of the preferable aspects of feedback (such as solution) could contribute to revision of writing in a near-anonymous public review environment. The results challenge us to probe deeper in order to identify strategies that will achieve substantial positive impact in practice.

Theoretical foundations and hypotheses

An increasing number of interventions are being designed and developed for peer grading in MOOCs (Kulkarni et al., 2015; Hicks et al., 2016; Joyner et al., 2016), ranging from novel grade aggregation techniques (Sajjadi et al., 2016), to enabling students to rate their graders in return (Staubitz et al., 2016). Prior work on peer assessment in MOOCs mostly focused on getting an accurate grade to students. The value for learning and improvement of performance as a result of feedback has been far less of a focus. Some work on infrastructure for supporting the feedback process in online settings provides a practical foundation for our work. For example, PeerStudio (Kulkarni et al., 2015) enables students to provide feedback to their peers in MOOCs. A class survey shows that students found free-form comments to be more useful than the grade, and students liked reading each other's work the most, more than getting feedback and revising their work. It was found in Joyner et al. (2016) that meta-reviewers wrote higher quality feedback based on the first round of reviews. In Tinapple et al. (2013), some comments from students showed how reviewing their peers' work increased accountability and a sense of community. For example, "You take the work into account much more when you are aware that your peers will be reviewing it" and "This class had a sense of community which I think many college courses lack. When someone feels more comfortable in the class, they will do better in that class." Staubitz et al. (2016) found that students had increased motivation to provide high quality reviews when they had the opportunity to rate their reviewers.

Such prior work has shown the potential for positive impact when giving students access to more reviews. Our work points to the accountability issue where students may feel more obliged to provide high quality feedback when their feedback is disclosed to the class instead of provided to individuals privately. On the other side, prior work has also shown that a lack of sense of community could give rise to feelings of disconnection and affect students' persistence (Kerka, 1996), while emphasizing community reduces drop out (Tinto, 1993). Exchanging feedback in a discussion forum may provide this needed sense of community.

There has been a line of research in the Learning Sciences that studies what makes feedback effective. For example, an authoritative reference is a coding manual for feedback quality developed in widely acknowledged work of Nelson & Schunn (2009). This coding manual was informed by a survey of prior work on what kinds of feedback led to better performance outcomes. For example, Ferris (1997) found that feedback that included summaries promoted more substantive responses to feedback. The same work also reported that specific comments were more helpful than general comments. Bitchener et al. (2005) found that feedback that contained solutions was more helpful for adults' writing performance. We use the framework contributed by this body of research as a basis for both scaffolding and assessment of feedback in our study.

Based on this review of prior work, we propose the following two hypotheses. ***(1) We hypothesize that feedback generated in a public environment that includes scaffolding for offering effective feedback will be of higher quality than feedback generated in an otherwise equivalent private environment.*** In addition to the effect on feedback quality, we will also examine whether feedback received in the public environment actually helps students improve their writing. ***(2) We hypothesize that students who receive feedback in a public environment will have a higher revision rate and show a higher quality of revision than students who receive feedback in a private environment.***

Method

We tested our hypotheses in a high internal validity setting as preparation for later research in high external validity large-scale online learning settings. We adopted a between-subjects design with random assignment in which we manipulate whether students write peer review in a public discussion forum or privately on a web page. The study was run as a lab study in an online crowdsourcing environment, as we describe below. In both conditions, participants complete a 4-step task in which they write individual assignments and provide feedback to one other student's assignment. In order to measure the quality of student feedback, we adapted the coding manual used in Nelson & Schunn (2009) and Patchan et al. (2016), and manually coded the feedback generated in the experiment into 9 constructs. We also developed a reliable coding manual to operationalize the quality of student assignments as control variables in our subsequent analyses. We use the same coding manual to measure the quality of student revisions. In the following subsections, we will provide a detailed description of the task, the coding manual for feedback, and the coding manual for assignment and revision quality.

Task description

All participants in the study participated in a 4-step task framed as a unit in an environmental sciences course, which took approximately an hour. In step 1 (~10 mins), students were asked to read learning materials on 4 types of energy. In step 2 (~20 mins), students were asked to write an energy proposal for a city. In step 3 (~15 mins), students were asked to provide a review to one other student's proposal based on the scoring rubrics provided to them. In step 4 (~10 mins), students received feedback on their proposal, rated the feedback on how helpful they perceived it to be on a scale of 1-5, and then were offered the option of revising their own proposal.

The experimental manipulation took place in step 3. In the private condition, students wrote their review on a webpage, whereas in the public condition, students were directed to a discussion forum and wrote the review using a similar interface and the same instructions but housed within the forum that was open to all participants. In both conditions, we introduced a feedback prompt. The feedback prompt works as a rubric that asks students to evaluate the quality of assignments from three perspectives, namely, 1) Clarity of arguments, 2) Supporting evidence, and 3) Adequacy of arguments as shown in Figure 1. Each piece of feedback thus contains three *Comment Segment* corresponding to the rubric. In the feedback prompt, we also provide guidance to students on how to write high-quality feedback, by specifically pointing to the preferable constructs of feedback as studied in prior work, such as "please summarize", "be specific", "give potential fixes", etc.

Hi, welcome! Please provide feedback to your peer's assignment using the following rubric. Keep the two general guidelines in mind. Provide specific comments: point to exact places that were problematic; give examples, etc.
Try to be helpful to your peers, think about how to help them improve instead of punishing them for mistakes.
1. Clarity of arguments: Please check whether the writing flows smoothly so you can follow the main argument. In your comments, please first summarize what you perceived as the main points being made so that the writer can see whether the readers can follow the paper's arguments. Then make specific comments about what problems you had in understanding the arguments and following the flow across arguments.
2. Supporting evidence from the readings: Please check whether the author's arguments are supported by evidence. If you found points made without support, describe which ones they were. If the support provided doesn't make logical sense, explain what that is. If some obvious counterargument was not considered, explain what that counterargument is. Then give potential fixes to these problems if you can think of any.
3. Adequacy of arguments: Please check whether enough perspectives or tradeoffs are considered. If you can think of any other factors that could play a role in a city's energy running, it is always a great idea to include different perspectives and provide extra information.

Figure 1. Feedback Prompt

Feedback quality coding

Our unit of analysis for feedback coding is *Comment Segment*. We adapted the first version of our coding manual from Nelson & Schunn (2009). In addition to the existing feedback constructs, we added *Localized Praise*. Though prior work has shown that *Praise* in feedback is at best ineffective and often diminishes future performance (Kluger & DeNisi, 1996), we might expect localized praise to be more helpful in preserving confidence than general praise. We also introduced a distinction in specificity for *Problem* and *Solution*, informed by Patchan et al. (2016), to distinguish feedback that focused on literal language usage (at a writing specificity level) and feedback that focused on logic or arguments (at an idea specificity level). Two researchers iterated on the definitions in the coding manual until sufficient inter-rater reliability was achieved over unseen data. In the final iteration, inter-rater agreement was evaluated for each category on 27 *Comment Segment* with final Kappa displayed in Figure 2. Once agreement was established, one of the researchers coded the whole set of feedback, blind to condition. A brief version of our coding manual is shown in Figure 2. The categories in the coding manual are not mutually exclusive; a comment could fall under multiple categories.

Subjective rating of feedback

Students are asked to rate how helpful the feedback comment is on a 1-5 scale, with 5 being "Very helpful" and 1 being "Very unhelpful".

Code	Definition	Example	Kappa
Summary	A list of the topics discussed in the paper, a description of the claims the author was trying to make, or statements of an action taken by the writer.	"The main point that I took away from this recommendation is that Plan B would be the best option due to the decrease in environmental risks associated with Wind Power."	0.64
Problem	A comment that describes what is wrong with the paper. First need to decide whether the problem is localized or not.	"Overall there was a lack of coherence and I don't feel there is a substantial enough argument here."	0.92
Problem Localized Writing	A comment that addresses an issue dealing with the literal writing, usually at a word level.	"There were too many 'and' words that made it very difficult to follow what the writer was saying in the proposal."	0.78
Problem Localized Idea	A comment that addresses an issue with the accuracy or completeness of the content in the paper.	"The only problem I see is the fact that having wind turbines and the main source of energy will require quite a large amount of wind turbines."	0.91
Solution	At least one solution or suggestion is explicitly offered. The solution should contain constructive insights, not rewrite the problem. First need to decide whether the problem is localized or not.	"I would only suggest maybe going more in depth on why these other alternatives are more useful."	0.79
Solution Localized Writing	A comment that addresses an issue dealing with the literal writing, usually at a word level.	"I think it would be easier for me to get into your proposal if you stated that you're proposing Plan 2 in the beginning of your argument. Then, giving us reasons why the other plans are not the best would make an argument that is easy to follow."	0.65
Solution Localized Idea	A comment that addresses an issue with the accuracy or completeness of the content in the paper.	"I think that a solution to this could be to emphasis their strengths and maybe sell some of there energy from nuclear plants to adjacent towns to make up the difference."	0.71
Praise	Complimentary comment or identifying a positive feature or talking about personal feelings/experience in the paper	"The proposal is clear, concise and very well written."	0.81
Praise Localized	At least one positive feature or good practice can be easily located by public viewers	"Plenty of factors and perspectives were added the tradeoffs were also considered. E.g. tourism, safety and lack of waste."	0.74

Figure 2. Coding manual for feedback

Assignment quality coding

Students were asked to review assignments based on a rubric that focuses on three distinct aspects of the assignment: 1) Clarity of arguments, 2) Supporting evidence, and 3) Adequacy of arguments. We evaluated the assignments from the same three aspects. For clarity of arguments, we drew from the writing rubric from Purdue University (2017) and assigned scores of 1-4 to represent the clarity level (1: Beginning, 2: Developing, 3: Proficient, 4: Mastery). To evaluate the quality of supporting evidence, we counted the number of factors (knowledge points) students mentioned in their proposal, from a list of factors including budget, waste disposal, carbon emission, tax credit, wildlife protection, water quality, etc. If the author wrote arguments that were contrary to the information provided in the learning materials, *e.g.*, "Coal energy is environmentally friendly," we counted that as an argument without supporting evidence. Using the above standards, we operationalized the extent of supporting evidence in student writing in two categories, number of *Arguments* mentioned, and number of *Unsupported Arguments*. To evaluate the adequacy of arguments, we counted the number of tradeoffs or comparisons that the author has mentioned in the writing.

Using this coding approach, we formalized the evaluation of each assignment into 4 numeric variables, namely: *Receiver Clarity*: a score in the range 1-4; *Receiver Number of Arguments*: a score in the range 1-11 in our data; *Receiver Number of Unsupported Arguments*: a score in the range 0-3 in our data; and *Receiver Number of Comparison/Tradeoffs*: a score in the range 0-5 in our data.

After students had the opportunity to make revisions as part of the task, we computed the difference in their text between their original proposal and the revised proposal. We then used the same coding manual to code the quality of revisions. If the revision addressed a clarity issue, we coded revision clarity as 1, otherwise 0. We also coded the number of arguments, unsupported arguments and tradeoff / comparisons in the revision.

Participant recruitment

We ran the study as a lab study online through Amazon Mechanical Turk crowdsourcing platform from October to November in 2017. We ran the experiment in batches, with each batch associated with one or the other condition of the peer review environment. Therefore, each batch was assigned either to the public peer review condition or the private peer review condition. We restricted the participants to have above 98% acceptance rate in their work history and be located in the US. We did not collect information about participants' demographics and education history, since their prior knowledge on the task would be accounted for by their assignment quality. The on-task time was ~55 minutes for each participant, with a possible 15-minute wait period. We compensated each participant $6 for their participation. We will further discuss the generalizability from crowdsourcing platforms to real MOOCs in the discussion section. For the public condition, we pre-populated the discussion forum with some existing student assignments. There were 3-7 people in each batch. They were

matched to provide feedback to each other's assignments. In the end, we recruited 65 participants, with 32 in the private condition, and 33 in the public condition. We only included participants who completed the entire task. Though participants were told to write review to only one student, 3 students in the public condition wrote one more review voluntarily. This resulted in a total number of 68 pieces of feedback, with 204 *Comment Segments*.

Results

Hypothesis 1A: Investigating differences in feedback quality

In order to test our first hypothesis, we operationalized feedback quality in two ways, expert coding and subjective rating as introduced in the methods section. In this section, we present results regarding the effect of our intervention on the quality of feedback indicated by expert coding. We built regression models to compare feedback quality between the public and private conditions. Since we operationalized feedback quality into 9 constructs, we built 9 models to compare the quality difference for each feedback construct respectively.

For each feedback construct, we set up a regression model to compare its frequency between the public and private conditions. In each model, the dependent variable is the feedback construct (*e.g.*, *Solution*), and independent variables include *Comment Segment*, *Condition*, and the *Interaction* between the two, as we anticipated the condition might have different effects in the three comment segments due to their different focus. We also introduced control variables indicating the quality of the original assignment submission, with the understanding that it would be harder for the reviewers to point out problems or solutions if the assignment was of high quality. As we introduced in the methods section, we operationalized assignment quality into 4 constructs, *Receiver Clarity*, *Receiver Arguments*, *Receiver Arguments Unsupported*, and *Receiver Tradeoff*. We nested each of the assignment quality indicators within *Comment Segment* in the regression models.

From the 9 regression models, we found that feedback generated in the public condition showed significantly higher numbers of *Solution* statements than feedback generated in the private condition, with $F(1, 186) = 4.765$, $p < 0.05$. The effect size value computed by Cohen's D is 0.71, suggesting a moderate to high effect. We also found that when the assignment included more tradeoff points, there were fewer solutions pointed out in the feedback, with $F(3, 192) = 3.369$, $p < 0.05$. This demonstrates that when the assignment is of higher quality, it was harder for reviewers to propose solutions. In addition, we also saw a significant difference in the frequency of *Solution Localized Idea* between the two conditions, with $F(1, 186) = 3.977$, and $p < 0.05$, with an advantage to the public condition. The effect size value computed by Cohen's D is 0.71, suggesting a moderate to high practical significance. This suggests that feedback generated in the public condition elicited more localized solutions that target at ideas.

We also found when student assignments showed a higher number of unsupported arguments and a lower number of tradeoffs, there tended to be more localized solution statements about ideas pointed out in the feedback, regardless of condition. The effect of number of unsupported arguments is marginally significant, with $F(3, 186) = 2.242$, $p = 0.085$ and the effect of number of tradeoffs is significant, with $F(3, 186) = 2.839$, $p < 0.05$. This again suggests that students tend to show more specific and substantive feedback when the assignment quality is lower. We do not see a significant difference in the frequency of other constructs of feedback between the two conditions. The average and standard deviation of the frequency for each feedback construct (*i.e.*, count of segments that include that construct) is displayed in Table 1. These per construct descriptive statistics are offered to provide an overview of the distribution of feedback constructs between conditions, not meant as a formal between-condition comparison.

Table 1: Descriptive statistics of the 9 feedback constructs by condition

Feedback Construct	**Private**		**Public**	
	Mean	**Std.**	**Mean**	**Std.**
Summary	0.43	0.50	0.39	0.49
Praise	0.57	0.50	0.65	0.48
Praise localized	0.15	0.36	0.13	0.33
Problem	0.53	0.50	0.45	0.50
Problem localized writing	0.06	0.24	0.04	0.19
Problem localized idea	0.4	0.49	0.39	0.49
Solution*	0.35	0.48	0.44	0.50
Solution localized writing	0.07	0.26	0.06	0.23
Solution localized idea*	0.27	0.45	0.34	0.48

Hypothesis 1B: Investigating differences in perceived quality of feedback

In this section, we present results regarding the effect of our intervention on the quality of feedback indicated by student subjective rating. We built a regression model using the *Subjective Rating* students gave as the dependent variable and the *Condition* as the independent variable. We included the assignment quality indicators as control variables, operationalized by the four constructs of assignment quality discussed above. Note that the rating was assigned to the feedback as a whole rather than each comment segment separately.

In the regression model, we found that feedback generated in the public condition received a significantly higher rating than in the private condition, with $F(1, 63) = 7.238$ and $p < 0.01$. The effect size value computed by Cohen's D is 3.06, suggesting a very high practical significance. The number of *Unsupported Arguments* in the assignment predicted lower rating of the feedback by the receiver with marginal significance, with $F(1, 63) = 3.65$, $p = 0.06$.

In Hypothesis 1A, we found that our intervention successfully increased the presence of solution and localized solution at an idea specificity level in student feedback. We want to further investigate whether the higher subjective rating in the public condition came from the elevated presence of these two constructs. We added the 9 feedback constructs as independent variables into the baseline regression model as described above. In the new model, we see that *Condition* and *Unsupported Arguments* still significantly predict higher feedback ratings, with $F(1, 189) = 18.3$, $p < 0.0001$, and $F(1, 189) = 11.44$, $p < 0.005$ respectively. Among the 9 feedback constructs, only *Problem Localized Idea* is marginally significant in predicting higher rating, with $F(1, 189) = 2.77$ and $p = 0.097$. We do see students in the public condition liked their feedback better, though we do not see the effect coming from the constructs of feedback that our intervention manipulates, namely *Solution* and *Solution Localized Idea*. Through the above analyses, we confirm that hypothesis (1) is generally supported. Feedback generated in a public environment shows signs of being higher quality than feedback generated in a private environment, indicated by both expert coding of feedback quality and student subjective rating.

Hypothesis 2: Investigating the effect of condition on revision of writing

In order to test our second hypothesis, we first built a regression model to compare the revision rate between the two conditions. Though we found that there was a trend that revision rate in the public condition was higher, we did not see a significant difference between the two conditions in students' revision rate, with $F(1, 63) = 0.22$, $p = 0.64$. This suggests that the aspects of high quality feedback manipulated through elevated accountability in our intervention may not be the same ones that are most important for achieving impact on writing. In order to understand this more deeply we conducted a post-hoc analysis.

In particular, we conducted a correlational analysis to investigate what constructs of feedback students receive are associated with increases in propensity to revise. We used the binary variable indicating whether students revised or not as the dependent variable and 4 constructs of assignment quality as control variables. We included the 9 constructs of feedback quality as independent variables. We found that the presence of *Problem* and *Solution* in feedback significantly predicts whether students revise or not, with $F(1, 190) = 6.13$, $p < 0.05$, estimate of coefficient = 0.3, and $F(1, 190) = 5.00$, $p < 0.05$, estimate of coefficient = 0.3 respectively. This is consistent with prior work that shows the presence of *Problem* and *Solution* is beneficial for student implementation of the feedback. However, we also found the presence of *Summary* in feedback negatively predicts whether students revise or not, with $F(1, 190) = 3.93$, and $p < 0.05$, estimate of coefficient = 0.13. This is inconsistent with prior work that found summary to be helpful for implementation. From a follow-up correlation analysis, we found that the presence of Summary is negatively correlated with the presence of Problem and Solution, and positively correlated with the presence of Praise. This suggests that when students are spending effort summarizing what they've seen in the proposal, they are less likely to spend time pointing out specific problems and solutions in the assignments.

In addition to students' propensity to revise, we also investigated the revision quality differences between the two conditions. In each of the regression models, we used one of the four constructs of revision quality as the dependent variable (*e.g.*, *Revision Tradeoff*), and used *Condition*, *Occurrence of Revision*, and the four constructs of assignment quality as control variables. We found the public condition showed more tradeoffs in revisions than the private condition; the effect is marginally significant, with $F(1, 63) = 3.159$, and $p = 0.08$. We do not see a difference in any other revision quality constructs.

Through the above analyses, we confirm that hypothesis (2) is only partially supported at best. There is no main effect of condition on propensity to revise between the public and private conditions. There is only a trend showing students in the public condition made higher quality revisions than students in the private condition when they did revise, and the difference is only marginally significant.

Discussion and future work

Public peer review intervenes on certain dimensions of feedback

Our experiment found that the manipulation of perceived supervision through a public peer review environment increased the presence of general solution and localized solution at an idea specificity level in student feedback. On the positive side, *Solution* is indicated in prior work to be the most important aspects that influence feedback quality (Nelson & Schuun, 2009; Patchan, 2016). On the negative side, more work needs to be done to achieve more substantial effects on improving feedback quality. Since the intervention of public peer review environment aims at manipulating students' accountability in feedback writing, the result might also suggest that elevating effort to offer feedback is not the same as increasing the ability to offer feedback. The public environment and feedback prompt offers increased awareness of the aspects that are desirable in high quality feedback, but more or different support may be required to address those aspects of feedback that students lack the skill to offer. The result provided in this experiment suggests that providing a public venue for feedback is one step in the right direction, but more is needed to achieve the ultimate goal of high quality feedback.

From feedback to revision

In our experiment, we observed that the public condition increased the presence of *Solution* and *Solution Localized Idea* in student feedback and also made students like their feedback better. However, it is surprising to see that increasing the prevalence of characteristics of feedback shown in earlier work to be associated with revision of writing (Nelson & Schunn, 2009; Patchan, 2016) did not show strong effects here. There are multiple explanations of this finding. First, it might be the case that it is not enough to offer more solutions, but qualities of the solutions offered are also important. It again points to the idea that raising awareness of what component needs to be included in feedback is not enough, rather future work needs to explore how to increase students' skill at offering feedback. Second, it could be that students lacked sufficient support on appropriation of feedback in their revision process. (Wichmann et al., 2017) This suggests that it could be fruitful to provide explicit instruction and support to students on how to incorporate feedback to revise their writing.

Implications for peer feedback research

Our experiment displayed some consistent findings with prior work in different educational settings that reinforced known mechanisms on feedback offering. For example, when the assignment is of lower quality, it is easier for students to offer solutions in their feedback; when the feedback contains *Problem* and *Solution*, the feedback receivers showed higher propensity to revise. However, it is inconsistent with prior work that when the feedback contains *Summary*, the receiver is less likely to revise. This could be explained by the fact that when students spent more effort writing *Summary* in the feedback, they spent less effort writing *Problem* and *Solution*. We consider this result to be contextual to the kinds of learning environment, where students spend minimal effort on the task. We also found that student subjective rating of feedback is not correlated with expert coding. Researchers in the future should thus be cautious of choosing metrics to evaluate feedback quality.

We acknowledge the limitations of this work as a lab study. Since the factor manipulated is related to student accountability, we must acknowledge that this audience might not demonstrate the expected effect on accountability as we would see in a higher-stakes learning environment. However, the advantage of a lab study is high internal validity, especially experimenting with factors that are not easy to directly manipulate in a real course. Coetzee et al. (2015) pointed out that though participants from crowdsourcing platforms may likely have different motivations from MOOC learners, their remote individual work setting without peer contact resembles today's MOOC setting where most students learn in isolation. Prior work (Wen et al., 2016; Wang et al., 2017) also demonstrated the potential of such high internal validity experiments in informing subsequent high external validity deployments in real MOOCs, which will be an immediate next step of our work.

Conclusion

In this study, we investigated the effectiveness and potential of a new peer review paradigm—public peer review—where students get full exposure to assignments and reviews of the class. We hypothesized that the increased accountability of the public environment would result in higher quality feedback being exchanged. We presented the results of an experimental study comparing feedback quality and revision rate between the public and private peer review conditions. The results support our hypotheses that students in the public condition provided more solutions and localized solutions about substantive ideas in the feedback they gave, and also perceived the feedback they received to be better than students in the private condition. Though we did

not observe a difference in students' propensity to revise their assignments between the two conditions, we see a trend showing students in the public condition demonstrated higher revision quality.

References

Bitchener, J., Young, S., Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, 14, 191–205.

Chi, M. T., Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219-243.

Coetzee, D., Lim, S., Fox, A., Hartmann, B., Hearst, M. A. (2015). Structuring interactions for large-scale synchronous peer learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1139-1152). ACM.

Ferris, D. R. (1997). The influence of teacher commentary on student revision. *TESOL Quarterly*, 31, 315–339.

Graham, S., Hebert, M., Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, 115(4), 523-547.

Hicks, C. M., Pandey, V., Fraser, C. A., Klemmer, S. (2016). Framing feedback: Choosing review environment features that support high quality peer assessment. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 458-469). ACM.

Joyner, D. A., Ashby, W., Irish, L., Lam, Y., Langson, J., Lupiani, I., Bruckman, A. (2016). Graders as Meta-Reviewers: Simultaneously Scaling and Improving Expert Evaluation for Large Online Classrooms. In *Proceedings of the Third ACM Conference on Learning@ Scale* (pp. 399-408).

Kerka, S. (1996). Distance Learning, the Internet, and the World Wide Web. *ERIC Digest*.

Kluger, A.N. and DeNisi, A. (1996) The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, 119(2):254.

Koedinger, K., Corbett, A. T., Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science practice chasm to enhance robust student learning. *Cognitive science*, 36(5), 757-798.

Kulkarni, C. E., Bernstein, M. S., Klemmer, S. R. (2015). PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the Second ACM Conference on Learning@ Scale*.

Nelson, M. M., Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science*, 37(4), 375-401.

Patchan, M. M., Schunn, C. D., Correnti, R. J. (2016). The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational* Psychology, 108(8), 1098.

Purdue University, College of Science writing rubrics (retrieved on July, 2017) https://www.science.purdue.edu/Current_Students/curriculum_and_degree_requirements/writing_rubric_gray.pdf

Sajjadi, M. S., Alamgir, M., von Luxburg, U. (2016). Peer Grading in a Course on Algorithms and Data Structures: Machine Learning Algorithms do not Improve over Simple Baselines. In *Proceedings of the Third ACM Conference on Learning@ Scale.*

Staubitz, T., Petrick, D., Bauer, M., Renz, J., Meinel, C. (2016). Improving the Peer Assessment Experience on MOOC Platforms. In *Proceedings of the Third ACM Conference on Learning@ Scale*.

Suen, H. K. (2014). Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning*, 15(3).

Tinapple, D., Olson, L., Sadauskas, J. (2013). CritViz: Web-based software supporting peer critique in large creative classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology*, 15(1), 29.

Tinto, V. (1993). Leaving college rethinking the causes and cures of student attrition: University of Chicago.

Wang, X., Wen, M., Rosé, C. (2016). Towards Triggering Higher-order Thinking Behaviors in MOOCs. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge.*

Wang, X., Wen, M., Rosé, C. (2017). Contrasting Explicit and Implicit Support for Transactive Exchange in Team Oriented Project Based Learning. In *Proceedings of the 12th International Conference on Computer Supported Collaborative Learning* (pp. 25-32). International Society of Learning Sciences.

Wen, M., Maki, K., Dow, S., Herbsleb, J., Rosé, C. (2018). Supporting Virtual Team Formation through Community-Wide Deliberation. In *Proceedings of the 21st ACM Conference on CSCW.*

Wichmann, A., Funk, A., Rummel, N. (2017). Leveraging the potential of peer feedback in an academic writing activity through sense-making support. *European Journal of Psychology of Education*, 1-20.

Acknowledgements

This work was funded by NSF Grants ACI 1443068, STEM+C/IIS 1546393. Thanks to all study participants for their time, and to Anhong Guo for his help and support.