Graders as Meta-Reviewers: Simultaneously Scaling and Improving Expert Evaluation for Large Online Classrooms

David A. Joyner[1,2], Wade Ashby[1], Liam Irish[1], Yeeling Lam[1], Jacob Langston[1], Isabel Lupiani[1], Mike Lustig[1], Paige Pettoruto[1], Dana Sheahen[1], Angela Smiley[1], Amy Bruckman[1], & Ashok Goel[1]

[1]Georgia Institute of Technology
School of Interactive Computing

[2]Udacity, Inc.

{david.joyner, mashby6, liam, yeeling.lam, jlangston9, isabel.lupiani, mikelustig, ppettoruto6, danasheahen, smiley}@gatech.edu; {asb, ashok.goel}@cc.gatech.edu

ABSTRACT

Large classes, both online and residential, typically demand many graders for evaluating students' written work. Some classes attempt to use autograding or peer grading, but these both present challenges to assigning grades at for-credit institutions, such as the difficulty of autograding to evaluate free-response answers and the lack of expert oversight in peer grading. In a large, online class at Georgia Tech in Summer 2015, we experimented with a new approach to grading: framing graders as meta-reviewers, charged with evaluating the original work in the context of peer reviews. To evaluate this approach, we conducted a pair of controlled experiments and a handful of qualitative analyses. We found that having access to peer reviews improves the perceived quality of feedback provided by graders without decreasing the graders' efficiency and with only a small influence on the grades assigned.

Author Keywords

Peer review; online education.

ACM Classification Keywords

K.3.2. Computer and Information Science Education.

INTRODUCTION

Recent trends in the development of higher education have introduced significant questions about scaling traditional university offerings. Massive open online courses (MOOCs) can draw tens of thousands of students each paying little to no money, making a traditional manual grading process using expert graders impossible. Many of these courses have opted to use peer grading to replace expert grading in assessment, and some studies suggest peer grading can be as reliable as expert grading under the right circumstances [6, 7, 21].

L@S 2016, April 25 - 26, 2016, Edinburgh, Scotland , United Kingdom.

ACM 978-1-4503-3726-7/16/04…$15.00 .
DOI: http://dx.doi.org/10.1145/2876034.2876044

However, increasing class sizes is not limited only to MOOCs; traditional educational offerings have begun experimenting with scaling up as well. The Georgia Tech Online Masters of Science in Computer Science (OMSCS) program is one such experiment. While neither Massive[1] nor Open[2], the program is still very large for a Master's program, with individual classes typically drawing between 200 and 500 enrollees. In the first two years of the program, there have been 11,000 individual course enrollments from 3,000 students. Although this large size is due in large part to the affordability (the entire Master's degree is around $6,600 [12], or $200 per month including fees assuming one class per semester), it also comes from an intentional decision by Georgia Tech to admit any qualified applicant rather than admit applicants based on the number of available seats.

These large class sizes, coupled with the drastically lower tuition cost, make scaling a key concern for educational programs such as Georgia Tech's OMSCS. However, given that these classes are often for credit for a Master's degree that is equivalent to the on-campus degree (rather than an implicitly weaker "online" degree), we argue that Georgia Tech cannot rely solely on peer grading to generate class grades. Part of the reputation of the Georgia Tech Master's degree is the trust that a graduate's work has been evaluated by experts. Heavy use of peer grading would compromise that reputation in the absence of systematic assurances of the equality of peer and expert grading.

Thus, the Georgia Tech OMSCS needs to scale up assignment assessment, yet it cannot use MOOCs' peer grading method for scaling! In this work, we investigate a novel compromise: could peer review be used to support, rather than supplant, expert grading? By equipping expert graders with peer reviews, we aim to make graders more like meta-reviewers in the traditional academic peer review process. Graders still assign the ultimate grade on every

[1] In its second year, the program has nearly 3000 students, making it the largest graduate program on campus [8], but much smaller than many MOOCs [13]. The program also hires TAs at a rate proportional to student enrollment.

[2] Participants in the OMSCS program must apply and be accepted to Georgia Tech.

assignment, but the feedback written by graders is informed by the feedback students received during peer review. We hypothesized that graders, now acting as meta-reviewers, may help students focus on the most important points from the peer reviews, resolve disagreements between peer reviews, and leverage peer reviews in writing their own feedback for the student. More importantly for scale, graders may ensure all students receive adequate feedback without spending additional time on grading: they may simply confirm strong peer reviews left by some students and devote more time to providing quality feedback to students who did not receive strong peer reviews.

To investigate the effects of reframing expert grading as meta-reviewing, we set up a pair of controlled experiments as well as gathered considerable additional data. We wanted to investigate several variables potentially affected by treating graders as meta-reviewers, including the quality of written feedback graders gave to students, the efficiency of the grading process, and the ultimate grades assigned by graders. We also wanted to investigate several interesting effects present in this data, such as the overall correlation between peer-assigned and expert-assigned grades, the potential biasing effect of seeing peer-assigned grades during expert grading, and the subjective impressions of graders on changing their role to meta-reviewers.

RELATED WORK

We want to contrast peer-to-peer feedback with peer assessment. In peer-to-peer feedback, students evaluate and give feedback on one another's work, but their evaluation does not affect the recipient's grade. In peer assessment, the peer's assigned grade is actually used in the calculation of the recipient's class grade. These terms are often used interchangeably in the community, but here, we will strictly use 'peer assessment' to refer to workflows wherein peer reviewers have influence over the recipient's grade.

Peer-to-peer feedback is well-documented to have a strong positive effect on learning [4, 19, 27] in a variety of contexts, including second language learning [23], college-level writing [31], and high school computing [28]. It is sometimes assumed that the main learning that occurs during peer-to-peer feedback is from the feedback one receives from one's peers; however, the literature and our experience both tell us that one of the most significant benefits of peer-to-peer feedback is in giving feedback [20].

Peer assessment has also been researched thoroughly. Existing studies suggest that it is possible to design assessments and associated grading instructions that will lead to peer-assigned grades being as reliable as expert-assigned grades [7]; however, these studies also point out that this reliability is not guaranteed, and that there are numerous instances where peer-assigned grades are not reliable replacements for expert-assigned grades. Notably, peer-assigned grades were found to be more valid replacements for expert-assigned grades in science and engineering and in more advanced courses [6]. Specifically applicable to the scale of this program, research has been generally positive on peer grading in MOOCs [21], and significant efforts have been devoted to increasing the reliability of peer grading specifically for this purpose [18, 25, 30], including through machine learning [24].

Whether or not the grades generated by peers are reliably similar to grades generated by experts is only one factor worth considering, however. Student perception is also an important factor. A recent study indicated that reliance on peer grading is one of the top drivers of high MOOC dropout rates [22]. This problem may be addressed by reintroducing some expert grading where possible.

BACKGROUND

This experiment took place in the Summer 2015 CS7637: Knowledge-Based Artificial Intelligence course [9, 10, 11] at Georgia Tech as part of the Georgia Tech's OMSCS program. 400 students enrolled in the 12-week summer offering of the course; 370 remained through the end of the initial drop/add week, and 280 completed the class.

The Course

CS7637 is a graduate-level course on the cognitive systems side of artificial intelligence. It is broken up into units on knowledge structures, learning, and reasoning, including analogical reasoning, metacognitive reasoning, and case-based reasoning. Originally offered in the OMSCS program in Fall 2014, it has become one of the program's highest-rated courses, earning its instructor and teaching assistant institute awards for the Fall 2014 offering.

In Summer 2015, CS7637 was offered as a 12-week course. As an online course, students do not attend synchronous lectures. Instead, lecture videos were produced and provided in partnership with Udacity. All videos were provided at the start of the course, along with a calendar of recommended viewing dates. Links to all course videos [9] and Summer 2015 course materials [14] are available in the references.

In addition to the pre-produced lecture videos, students participated in class discussions on a Piazza online forum [26]. Students were responsible for ~8500 contributions to the class Piazza forum, including proposing over ~750 discussions on their own, suggesting an active student community. For their work in the course, students completed three assignments, three projects, an unproctored final exam, and roughly 30 peer reviews.

The Assignments

Two types of written assignment were required in CS7637 in Summer 2015: written assignments [15] and project reflections [16]. Links to the descriptions of these assignments given to students are in the references section. Briefly, the written assignments all asked students to write ~1000 words applying certain course concepts to specific real-world problems. The project reflections asked all students to write ~1500 words reflecting on the successes and failures of their project and the relationship between

their agent and human cognition. Both these assignments are open-ended, and although multi-point rubrics are given to evaluate the assignments, grader subjectivity does come into play (although we conjecture that grader subjectivity is the same across all assignments). All analyses below are based on the written assignments; no experiments were conducted on the project reflections.

We assigned six written assignments, and students chose three to complete in a paired fashion (e.g. each student chose to complete either assignment 1 or assignment 2, 3 or 4, and 5 or 6). All students completed three peer reviews on *each* of the six assignments, however, and so students received an average of six reviews per written assignment they completed while giving three reviews for each assignment. All students completed all three project reflections and completed three project reflection reviews, and thus students both gave and received an average of three peer reviews on project reflections. Projects were submitted between each pair of assignments.

	Online Summer '15	On-Campus Fall '14
% over years 25 old	88%	18%
% female	14%	24%
% previously obtaining a Master's degree	29%	6%
% previously obtaining a Doctoral degree	8%	0%
% international students*	13%	68%
% working full-time	90%	5%
Predicted hours per week spent on the course	11.2 (5.6)	7.9 (4.0)
Years of programming experience	10.6 (7.4)	5.3 (2.9)

Table 1. Demographic differences between online students in the Summer 2015 class and on-campus students in the Fall 2014 class.

Student Demographics

The nature of the online Master's program dictates that the student body of the class was significantly different from that of traditional on-campus programs [17]. Table 1 above compares the demographics of the Summer 2015 course with the Fall 2014 on-campus offering of the same course. Statistics marked with an asterisk were not collected per-course, and instead are derived from official Georgia Tech statistics [8] or other sources. Numbers in parentheses are standard deviations where appropriate.

As shown, students in the online program are significantly older, more educated, and more experienced. They are also, arguably, busier: most have full-time jobs, and anecdotally, most have families. Given that this experiment is built around peer reviews written by students, these different demographics may have a significant impact on the generalizability of the results observed here.

About the Graders

During the summer 2015 semester, the class employed ten teaching assistants, nine of whom participated in the assignment grading described here. All nine graders were current OMSCS students who had received an A in the class in either Fall 2014 or Spring 2015. All of these graders had previously received Bachelor's degrees, and two had previously received Masters degrees in other fields. Six were working full-time during the summer semester while working as a teaching assistant, while three were not. Four were taking other classes in the OMSCS program while working as a teaching assistant, while five were not. Altogether, four were both working full-time and taking other classes while working as teaching assistants, exemplifying the need for efficiency in grading duties. Additionally, three had formal teaching experience, one as an adjunct instructor at the undergraduate level, one as a teaching assistant and music teacher, and one as an undergraduate teaching assistant. Two also had prior experience with teaching online, one within a university and one through Coursera.

PEER FEEDBACK

To support this experiment in having graders act as meta-reviewers, we used a tool developed at Georgia Tech called Peer Feedback. Peer Feedback was originally developed to support peer-to-peer reviews in online and on-campus classes. Students enroll in a class section on Peer Feedback and can then either submit assignments directly through the tool or have their assignments ported over from a learning management system (in our case, Sakai [29]). After the assignment deadline passes, the teaching team runs the pairing process, where each student in the class is randomly assigned three classmates to review. These pairings are not bidirectional: students are not necessarily evaluated by the same classmates that they evaluate. Peer review is also not anonymous: feedback authors and recipients knew one another's identities to emphasize collaboration and discussion rather than numeric scores. In fact, one of the most common requests of students is the ability to more easily continue the conversation with their peer reviewers. Note that this structure differs from the "on-demand" pairing used by many online courses, where a student is assigned a classmate to review when they begin the peer review process (e.g. [21]).

In addition to the peer review function that gives Peer Feedback its name, the tool also comes equipped with a grading workflow. In the grading workflow, graders can be added to the class, and are assigned assignments to grade just as peers are assigned peers to evaluate. Graders in the workflow grade on the same rubric as peers. When available and enabled, graders can also view the peer

reviews that a student received while grading that student's assignment.

Peer Reviewing

Participation in peer review accounted for 15% of each student's grade. A 'peer review' comprises two pieces: scores on a scale of 0 to 5 in each of seven (for project reflections) or eight (for written assignments) rubric categories and written feedback. Peers and graders evaluate projects according to the same rubrics; for the remainder of this paper, we will discuss only the total scores assigned by peers and graders, not the individual rubric scores.

For the Summer 2015 section of CS7637, students were informed that they would be graded in part on the quality of the written feedback they provided in their peer reviews. Students were not given explicit instruction on how to write good peer reviews, but were provided with exemplars of high-quality peer reviews from past sections of the class. Anecdotally, noting that reviews would be judged for their quality in addition to their completion led to a significant uptick in review quality. A simple indicator of this is the increased average length of peer reviews, from a couple dozen words in past semesters to roughly 100 words in Summer 2015.

All assignments in the course were due on Sunday night; on Monday mornings, students received pairings for their peer reviews. Peer reviews were due by seven days following the original assignment due date. Peer reviews were made available to students the moment they were received from peer reviewers. After receiving a peer review, the recipient had the option of leaving 'meta-feedback' evaluating the feedback they received on a scale of 1 to 7, with 1 meaning the least helpful and 7 meaning the most helpful; individual labels were not attached to each number to avoid challenges with numerical analysis on otherwise ordinal data, although such challenges may still persist. Students could also leave written commentary. These ratings and commentary were only seen by the teaching team. Leaving meta-feedback was strictly voluntary, although students were encouraged to do so several times throughout the semester. Students were also informed of the way in which meta-feedback ratings would be used in order to encourage further participation. Although students providing meta-feedback may not be representative of the class as a whole, we do not anticipate a systematic bias toward one condition or the other unique to those students providing meta-feedback. In this paper, 'meta-feedback score' refers to the numeric component of meta-feedback.

Expert Grading

After the due date for peer reviews (that is, seven days after the original submission deadline), grading began. This delay was to allow the graders to have access to peer reviews when grading. At the beginning of the semester, graders underwent a training exercise where all graders independently evaluated assignments from previous sections and discussed their disagreements. During actual grading with peer reviews, graders were asked to read the peer review feedback students had received when considering their own review. Graders were given some basic suggestions on how peer reviews might help them compose their own feedback: they might emphasize strong feedback that was received in the peer reviews, help resolve conflicts or disagreements among peer reviews, and leverage peer review comments in the feedback that they themselves provided for the student. This last dynamic represents some strong potential for scaling assessment to larger classrooms: by opening the possibility for graders to 'crowd-source' written feedback and simply pick feedback from the available strong reviews, they can grade more quickly and focus more time on those students that did not receive quality peer reviews [2]. Importantly, graders were explicitly told they should not automatically agree with peer review feedback or scores.

Graders were given seven days to finish grading their allotted assignments. At the conclusion of the grading process, a simple normalizing function was applied to smooth inter-grader differences in assigned scores, and then grades were released, two weeks after the original assignment deadlines. As with peer review, students had the opportunity to leave meta-feedback for the TAs evaluating them. This meta-feedback was seen only by the course instructor. Students were encouraged to participate in meta-feedback, but were not required to do so.

EXPERIMENTAL DESIGN

The experimental design merged design-based research [1] and controlled experimentation. Each pair of assignments was treated as an individual controlled experiment investigating a particular treatment and its impact on notable dependent variables. The data from each experiment were then analyzed, and the results informed the structure and questions for the next experiment. The calendar of deliverables allowed a week for analysis of the previous week's experiments while project reflections were graded. No experiments were performed on project reflection grading.

The primary function of these experiments was to improve the class itself; we wanted to know whether treating grading as a process of meta-reviewing benefited students and/or improved grading efficiency. To ensure fairness during this semester, grades were normalized within each assignment and within, rather than across, treatments (although later analysis proved this was largely unnecessary). We also asked students to consent to user research as part of their participation in the class; only those students who consented to participate in this research are included in the data for this analysis. This was a large majority of the class.

Experiment 1

The first experiment asked the most fundamental question of the study: does equipping graders with peer reviews improve the grading process without compromising the integrity of the assigned grades?

Experiment 1 Design

To investigate this question, we conducted a controlled experiment. Each assignment was randomly assigned to either the control or experimental condition. In the control condition, graders evaluated assignments the way they have done in past semesters, without peer reviews. In the experimental condition, graders had access to peer reviews while evaluating assignments. Thus, the independent variable was access to peer reviews during grading. Three dependent variables were examined: time spent grading each assignment, grade assigned to each assignment, and meta-feedback score received based on each assignment.

Experiment 1 spanned two assignments, Assignment 1 and Assignment 2. In Assignment 1, all graders were assigned roughly 18 assignments to grade. Half the graders were randomly assigned to grade their 'control' assignments first, and half were assigned to grade their 'experimental' assignments first. In Assignment 2, all graders were assigned roughly 23 assignments to grade, and all were assigned to switch the order of evaluation (control first or experimental first) compared to Assignment 1. Thus, each grader graded 41 assignments between Assignment 1 and Assignment 2, with approximately an equal number in the control and experimental conditions while alternating which batch of assignments was graded first to correct for ordering effects. 343 total assignments were graded for Experiment 1, with one more experimental assignment than control assignment.

Experiment 1 Results

Experiment 1 revealed that having access to peer reviews during the grading process had no statistically significant effect on the grades that graders assigned and no consistent effect on the time graders spent grading each assignment. However, access to peer review did lead to a statistically significant increase in the meta-feedback score assigned by students to the graders.

Of the 343 assignments graded for Experiment 1, 90 received meta-feedback (26%); 42 control assignments and 48 experimental assignments received meta-feedback. The average meta-feedback score across all assignments was 5.84 (σ = 1.38). The average meta-feedback score for control assignments was 5.43 / 7.00 (σ = 1.80, n = 42), and the average meta-feedback score for experimental assignments was 6.21 / 7.00 (σ = 0.74). Testing for the difference of these means gives t = 2.75 ($p < 0.01$), demonstrating that assignments graded after seeing peer reviews received a statistically significantly higher average meta-feedback score than assignments graded without seeing peer reviews. The 95% confidence interval for the difference in meta-feedback score between the control and experimental conditions is 0.20 to 1.36. Described differently, students rated the feedback they received from graders as 11.1% better when the grader had access to peer reviews, with a 95% confidence interval of 3% to 19%.

The second dependent variable we investigated was grading efficiency. Of the nine graders, seven graded their assignments in such a way that we were able to glean reliable statistics from Peer Feedback about how long they spent grading their assignments. The graders varied significantly in the difference in the amount of time spent grading control and experimental assignments; three graders spent more time on average on control assignments, while four graders spent more time on average on experimental assignments. A t-test comparing the average time spent per assignment per grader found no statistically significant difference (t = 0.88, $p > 0.10$) in the amount of time a grader spent on assignments in one condition over the other.

The third dependent variable we investigated was grades assigned. We wanted to ensure that having access to peer reviews during grading did not systematically bias graders in favor of or against the students' work. Analysis found that assignments in the control group received an average grade (out of 40) of 28.81 (σ = 6.63), while assignments in the experimental group received an average grade of 28.38 (σ = 5.74). A t-test comparing these means showed t = 0.6517 ($p > 0.10$), providing no evidence to support the idea that assignments graded after seeing their peer reviews received systematically higher or lower grades.

In regards to these last two, it is important to note that absence of evidence is not evidence of absence; these analyses are not sufficient to prove that grading time or average grades assigned remain equal when equipped with peer review; proving equality would require different analysis. However, these tests do fail to find evidence to support the hypothesis that access to peer review during grading influences grading efficiency or average grades assigned.

An additional possible criticism of this analysis is that it tests only whether access to peer review makes graders more generous or strict on average, not whether access makes them converge onto the peer-assigned grades. This suggestion is examined further under Further Investigation.

Experiment 2

Due to the perceived positive effect of having graders act as meta-reviewers over peer reviews, we decided all assignments for the remainder of the semester would be conducted in this fashion to maximize student learning. Although we observed no statistically significant difference in the grades assigned to assignments in the control and experimental conditions, students expressed concern that seeing the rubric results from peer reviews would influence the grader's opinion of the paper. So, in Experiment 2, we decided to ask: does equipping graders with only the text of peer reviews preserve the benefits seen previously while removing a perceived source of grader bias?

Experiment 2 Design

To investigate this question, we conducted another controlled experiment. Each assignment was randomly assigned to either the control or experimental condition, without concern for the condition to which the corresponding student's assignment was assigned in Experiment 1. In the control condition, graders evaluated assignments the way they had in the experimental condition in the previous experiment, with access to both written peer reviews and peer-assigned rubric scores. In the experimental condition, the peer-assigned grades were hidden, and graders could only see the written feedback. Thus, the independent variable in Experiment 2 was access to the peer-assigned rubric scores. Two dependent variables were examined: grade assigned to each assignment and meta-feedback score received based on each assignment. Time spent grading each assignment was not examined in order to allow the graders to use more natural workflows (such as previewing all assignments before beginning to grade) rather than grade each assignment in one sitting.

Like Experiment 1, Experiment 2 spanned two assignments, Assignment 3 and Assignment 4. In Assignment 3, all graders were assigned roughly 12 assignments to grade. Half the graders were randomly assigned to grade their 'control' assignments first, and half were assigned to grade their 'experimental' assignments first. In Assignment 4, all graders were assigned roughly 21 assignments to grade, and all were assigned to switch the order of evaluation (control first or experimental first) compared to Assignment 3. Thus, each grader graded 33 assignments between Assignment 3 and Assignment 4, with approximately an equal number in the control and experimental conditions while alternating which batch of assignments was graded first to correct for ordering effects. Fewer submissions were observed in assignments 3 and 4 together due to class withdrawals. 296 total assignments were graded in assignments 3 and 4, with an equal number of assignments in each condition.

Experiment 2 Results

Experiment 2 found no evidence to support the hypothesis that access to peer-assigned grades affects the meta-feedback scores received or the grades assigned by graders compared to access to the textual peer reviews without the peer-assigned grades.

We first examined meta-feedback scores between the control and experimental groups. 87 students left meta-feedback on the grades and written feedback they received from graders. Feedback on assignments graded in the control group (with peer-assigned grades and written comments) received an average meta-feedback score of 5.60 / 7.00 (σ = 1.47; n = 47). Feedback on assignments graded in the experimental group (with peer review written comments, but without peer-assigned grades) received an average meta-feedback score of 5.95 / 7.00 (σ = 1.45; n = 40). A t-test comparing these gave t = 1.128 ($p > 0.10$), providing no evidence for the hypothesis that access to full peer reviews would affect meta-feedback scores compared to access to written comments only.

We then examined the grades assigned by graders to assignments in Experiment 2. Assignments graded in the control group received an average grade (out of 40) of 28.98 (σ = 6.40; n = 148). Assignments graded in the experimental group received an average grade of 29.32 (σ = 6.88; n = 148). A t-test comparing these gave t = 0.4461 ($p > 0.10$), providing no evidence for the hypothesis that access to full peer reviews systematically affected grades assigned to those assignments compared to grades assigned to papers graded without seeing the peer-assigned grades.

As with Experiment 1, the structure of this experiment is insufficient to categorically prove that access to rubrics has no effect on meta-feedback or assigned grades compared to grading without rubrics. However, this investigation gives no support for the hypothesis such a relationship exists.

Grading Workflow Evaluation

Experiments 1 and 2 found that having graders operate as meta-reviewers over peer-assigned reviews had a statistically significant positive effect on the quality of reviews students received as assessed by the assigned meta-feedback ratings, while not influencing the time spent grading assignments or the grades ultimately assigned by graders. Discussion of the results of these experiments revealed, however, that a possible notable factor of the grading process was the different workflows in which each grader graded. Thus, during grading of the final two assignments, we developed a typology of the grading workflows in which graders engaged. This followed a qualitative research design. Graders for this experiment were instructed to carefully document their grading process, including each time they opened an assignment, reviewed an assignment, read a peer review, issued a grade, and revised a grade. These nine accounts of the grading process were gathered and summarized.

Typology of Grading Workflows

Analysis of the self-reported grading processes revealed considerable overlap, but also notable differences. All graders reported beginning the grading process by first reviewing the recorded lectures for the topics covered by that assignment, as well as the directions for the assignment itself. From there, the workflow split: seven graders began grading, while two graders first took a cursory look over all their assignments to set their overall expectations before returning to the first assignment and beginning to grade. Each grader then reported participating in a roughly similar process for each individual assignment: they initially read over the peer reviews, then the paper itself. All graders reported writing comments and composing their rubric scores while reading the paper. At the conclusion, the workflow branched again: five graders reported a phase at the end of revisiting their earlier grades (or leaving open the option to do so) to modify them as necessary if their attitudes changed based on grading the subsequent essays.

Thus, two of the nine graders previewed all assignments initially; five of the nine graders left room to return and revise their grades at the conclusion of grading; and two graders graded each assignment in order with no previewing or revisiting. Interestingly, however, all graders participated in generally the same process of reviewing each assignment's peer reviews before reading and grading the assignment.

There are two main takeaways of this evaluation. First, that some element of internal normalization happened within the majority of graders: seven of nine graders built in the ability to revise their grades based on their experience grading other assignments. Second, the location of peer reviews in the evaluation workflow was consistent across all graders, even without explicit instructions. All graders reviewed peer reviews prior to the student's assignment, rather than reading and grading the assignment prior to reading the peer reviews.

FURTHER INVESTIGATION

The previous experiments addressed the core question of this research: these experiments suggest that equipping graders with peer reviews during the expert grading process improves the quality of feedback students receive without affecting the average grades assigned or the average time spent grading. However, none of these experiments addressed the possible connection between peer and expert grades. If, for example, peer grades were found to be highly correlated with expert grades, an argument could be made (as has been made in the past [6, 7, 21]) that peer grades could supplant expert grades as a reliable grading mechanism. There would also be arguments that our analysis above was not sufficient to rule out a biasing effect of exposure to peer review results, as well as that meta-feedback is not a useful metric as they are likely to be more positive in response to positive grades and negative in response to negative grades.

Expert Grades vs. Peer Grades

Initially, we wanted to determine if peer grades were good predictors of expert grades. During the semester, 1751 submissions received a combined total of 7287 peer reviews. A scatterplot showing the plot of expert grades (y) as a function of average peer review grades (x) is in Figure 1. Linear regression revealed an R^2 value of 0.3769, with a regression equation of $y = 0.8235x + 6.936$. This R^2 value suggests a moderately strong relationship, but one that leaves too many outliers on either side to support using peer grading in lieu of expert grading. A follow-up analysis calculating coefficients of determination solely within written assignments and project reflections found similar results; for written assignments alone ($n = 907$), $R^2 = 0.2836$ ($y = 0.8399x + 6.602$), and for project reflections alone ($n = 844$), $R^2 = 0.4608$ ($y = 0.7972x + 7.465$).

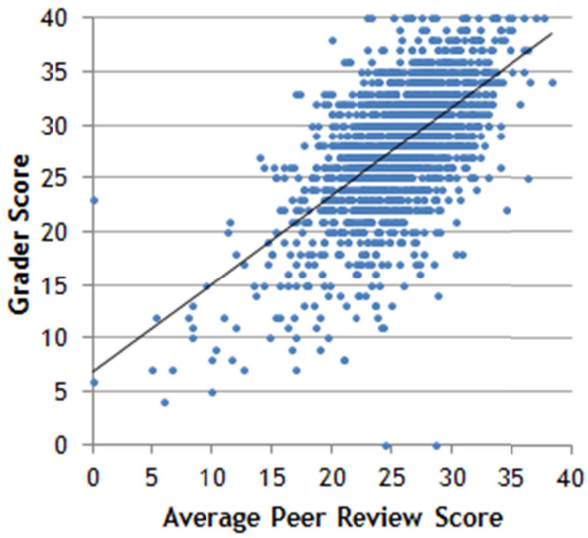

Figure 1: Average grader-assigned scores plotted against average peer review scores (n = 1751). The maximum grade was 40 points.

Bias Effects of Peer Reviews

The structures of Experiment 1 and Experiment 2 are not sufficient to completely rule out the hypothesis that access to peer review during grading biases graders to agree with the peer reviews. These analyses are only performed on the average grades assigned on all assignments. If biases existed in both directions within the experimental groups, they may cancel each other out, leading to no average difference in assigned scores despite the presence of differences in grades assigned to individual assignments.

To examine this, the above investigation comparing expert and peer grades was extended to compare the control and experimental groups within each experiment. If access to the peer-assigned rubric scores biases graders to agree with the peer reviews, we would expect to see a greater correlation between peer- and expert-assigned grades in the experimental group from Experiment 1 and the control group in Experiment 2 (that is, the two treatments where graders had access to peer-assigned rubric scores). We may also expect to see greater correlation in the control experimental group in Experiment 2, where although graders did not see rubric scores, they nonetheless saw written feedback that likely indicates the reviewer's attitude.

Table 2, below, provides the coefficients of determination (R^2) between average peer grades and expert grades within each experiment and condition.

These data suggest that a small biasing effect does exist when expert graders have access to the rubric scores assigned by peers during peer review. The highest coefficients of determination were seen for those groups of assignments that were graded with access to the rubric scores assigned during peer review. The effect of this bias was relatively small. The overall coefficient of

determination between peer grades and expert grades when peer-assigned grades were visible during expert grading was 0.3203 (that is, putting together all treatments where peer-assigned grades were visible), indicating that access to peer-assigned grades during expert grading raised the coefficient of determination between peer-assigned grades and expert grades by 0.06 in these experiments compared to grading with no peer reviews available. This represents an average difference of roughly a point on these assignments. A higher coefficient of determination between peer-assigned grades and expert grades was observed for project reflections ($R^2 = 0.4608$); we speculate, however, this is due to project reflection grades being higher and less varied overall. Most notably, this effect disappeared when peer-assigned grades were hidden during expert grading, while in Experiment 2, meta-feedback scores did not decrease when peer-assigned grades were hidden. Therefore, the maximum possible benefit can be observed by equipping expert graders with written peer reviews, but hiding peer-assigned grades; this removed the bias of seeing peer-assigned grades while preserving the higher feedback quality associated with meta-reviewing as assessed by meta-feedback scores.

	Coefficients of Determination (R^2)	*n*
Experiment 1, Control Condition	0.2616	171
Experiment 1, Experimental Condition	0.3264	172
Experiment 2, Control Condition	0.3412	148
Experiment 2, Experimental Condition	0.2028	148

Table 2. Coefficients of determination between peer-assigned graders and expert grades within each experimental condition. A small increase in correlation is seen in those conditions where graders could see the rubric scores assigned during peer review.

Expert Grades vs. Meta-feedback

During our analysis, we speculated that there may also be an interaction between the grades assigned by graders and the meta-feedback scores returned by students. Specifically, we conjectured that students may be more likely to give higher meta-feedback scores when they received good grades, and lower meta-feedback scores when they received bad grades. To look for this effect, we performed a linear regression between expert-assigned grades and meta-feedback scores. Because written assignments and project reflections were graded on different scales, we performed these two analyses separately. For written assignments, we found $R^2 = 0.0923$ ($n = 241$); for project reflections we found $R^2 = 0.0187$ ($n = 190$). Thus, we determined there was no correlation between expert-assigned grades and meta-feedback.

Grader Impressions

Separate from the effect of access to peer review during grading on assigned grades, grading efficiency, and meta-feedback scores, we were also interested most subjectively in how the graders perceived access to peer reviews to affect their grading process. At the conclusion of Experiment 1, we conducted a short survey of the graders who had graded assignments 1 and 2 alternating between seeing and not seeing peer reviews during grading. Eight of the nine graders completed this survey. Based on this survey, graders reported that access to peer review was useful, but that it did not make their jobs easier; rather, they felt it was useful in improving the quality of their feedback. This corresponds to the opinion given by students through the meta-feedback ratings. The data and survey responses agree that access to peer review did not improve grader speed or efficiency, but it did improve the quality of feedback that students receive.

OPEN ISSUES

A few open issues remain that affect the generalizability of the conclusions observed in these analyses. First, as mentioned originally, the student body in this class is significantly different from the student body even in typical graduate-level classes, let alone undergraduate or K-12 classes [17]. These students are older, more professionally experienced, more educated, and more dedicated (based on how many hours they predicted they would spend on the course) than typical Masters students. We conjecture that this leads to improved peer review quality. Fortunately for this class, this would improve the amount of learning students receive from the peer review process, but it also means the usefulness of peer reviews may not extend to other classes and programs. Research applying these principles to other student bodies would be necessary to ensure this generalization.

Second, just as the student body in this study was rather unique, so also the graders were unique – all graders in this study were themselves students in the OMSCS program. All nine graders involved in this study demonstrated a significantly higher level of commitment to improving the class than graders the instructor had seen in past sections. Less-dedicated graders may react differently to access to peer reviews during grading. We have two hypotheses in this area. First, we hypothesize that less-committed graders would experience the benefits of access to peer review in their efficiency rather than their review quality. Second, we hypothesize that the commitment of the graders in this experiment led to a higher baseline for meta-feedback scores, and that less-committed teams of graders may actually experience greater relative benefit due to a lower baseline. In other words, the quality of graders in this experiment left relatively little room for improvement based

on access to peer reviews; other teams may have more room for improvement. Additional research would be necessary as well to identify whether the effects seen here are universal to all graders or are a function in part of the unique qualifications and dedication of the graders participating here.

Third, as noted, access to peer reviews was shown to have a small biasing effect on the grades awarded by graders. This biasing effect should be resolved by hiding peer-assigned grades during expert grading. However, given that peer-assigned grades ought to be hidden during expert grading, should peers assign grades at all? Should peers instead focus only on providing written reviews instead of rubric scores? Research has already suggested that expert grades can undermine intrinsic motivation [5] and diminish the receipt of written feedback [3], and so focusing peer review on written reviews rather than peer-assigned grades could improve the learning that results from peer review while still allowing graders to act as meta-reviewers.

Last, this solution to scaling feedback while preserving expert evaluation assumes that a program like Georgia Tech's OMSCS cannot use peer assessment while maintaining its rigor and reputation. This assumption may not be accurate; after all, if peer-assigned grades *are* perfectly reliable duplicates of expert-assigned grades, then peer assessment is effectively as good as expert assessment. Research indicates this may be possible [6, 7], but considerable attention must be paid to designing assessments and accompanying grading workflows whose grades will retain their reliability when evaluated by peers. In this study, for instance, while peer and expert grades were somewhat correlated, the correlation was not sufficient to replace expert grades with peer grades. Machine learning may provide a mechanism to realize this, however, by seeding peer assessment activities with assignments with known expert evaluations, and using peers' assessments of those assignments to evaluate their individual reliability as graders overall. This method would also build a consistent account for why individual peer assessment grades are reliable, rather than relying simply on prior data indicating the reliability of peer assessment.

CONCLUSIONS

This study examines the benefits and drawbacks of structuring assignment grading as a process of meta-reviewing in an online graduate-level class. In traditional grading, graders see only the paper itself when evaluating and writing feedback. In a meta-review, a meta-reviewer sees the original submission as well as a collection of peer reviews. The meta-reviewer's feedback, then, is informed by the feedback written during peer review. The meta-reviewer can acknowledge important pieces of feedback the student has received, resolve disagreements amongst the peer reviews, and ensure that all students receive good feedback by focusing more time on those students who did not receive useful feedback during peer review.

Our experiments and analyses identified a notable benefit, a potential drawback, and an important recommendation for treating grading as meta-review. First, treating graders as meta-reviewers led to better feedback, as evaluated by the students. Students rated the feedback from graders acting as meta-reviewers as 11% better than feedback from graders grading more traditionally. There was no observed decrease in graders' efficiency nor any overall biasing effect on grades due to meta-reviewing. However, a small biasing effect was observed on individual assignments: there was a slightly greater correlation between peer-assigned grades and expert grades when expert graders could see the peer-assigned grades while grading. This small biasing effect disappeared while the higher meta-feedback ratings remained when peer-assigned grades were hidden but written peer reviews remained available to graders.

Therefore, this work concludes that treating grading as meta-reviewing has significant potential to help large programs scale by improving the quality of feedback students receive without compromising the efficiency of the grading process. Additional research would be necessary to ensure that the same effects observed here are observed at different levels of education and with different levels of commitment from graders. In future work, we will evaluate how to distribute the benefits of this approach across the target outcomes. While improving feedback quality without decreasing efficiency is a good outcome, it will be useful – especially as courses and programs scale up – to create improvements to efficiency as well. Specifically, we will examine whether or not different teams of graders, different guidelines in grading, or different evaluation workflows increase efficiency while preserving review quality.

ACKNOWLEDGMENTS

We are extraordinarily grateful to Joe Gonzales for developing the Peer Feedback tool and equipping it with the features necessary to perform this work. We are also grateful to Bryan Wiltgen, another teaching assistant during the class that did not participate in this grading, for his support. We are also grateful to the innumerable people at the Georgia Tech College of Computing, Udacity, Georgia Tech Professional Education, and AT&T for their support of this class and this program.

REFERENCES

1. Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, *13*(1), 1-14.
2. Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: a clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, *1*, 391-402.
3. Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working Inside the Black Box: Assessment for Learning in the Classroom. *Phi Delta Kappan*, *86*(1), 8-21.

4. Boud, D., Cohen, R., & Sampson, J. (2001). *Peer Learning in Higher Education: Learning from & with Each Other*. Psychology Press.
5. Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology*, *58*(1), 1-14.
6. Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, *59*(4), 395-430.
7. Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70*(3), 287-322.
8. Georgia Tech Fact Book. (2014). Graduate Enrollment by College, Ethnicity, & Gender [Web Site]. Retrieved from http://factbook.gatech.edu/admissions-and-enrollment/graduate-enrollment-by-college-ethnicity-gender-table-4-15/
9. Goel, A. & Joyner, D. (2014). CS7637: Knowledge-Based AI: Cognitive Systems [Online Course]. Retrieved from https://www.udacity.com/course/ud409
10. Goel, A. & Joyner, D. (2014). Putting Online learning and Learning Sciences Together [Video]. Presented to the Georgia Tech GVU Brown Bag. Retrieved from https://www.youtube.com/watch?v=N56ghCGmWWQ
11. Goel, A., & Joyner, D. (2016). An Experiment in Teaching Artificial Intelligence Online. In D Haynes (Ed.) *International Journal for the Scholarship of Technology-Enhanced Learning(1)* 1.
12. Grimmelmann, J. (2014). The Merchants of MOOCs. *Seton Hall L. Rev.*, *44*, i.
13. Hu, H. (2013). MOOC migration. *Diverse Issues in Higher Education*, *30*(4), 10.
14. Joyner, D. (2015). CS7637: Knowledge-Based AI Summer 2015 Class Documents [Google Folder]. Retrieved from http://bit.ly/1PRZTCb
15. Joyner, D. (2015). CS7637: Assignment 6 (Summer 2015) [Google Document]. Retrieved from http://bit.ly/1XAkdtb
16. Joyner, D. (2015). CS7637: Project 1 (Summer 2015) [Google Document]. Retrieved from http://bit.ly/1jSOyF6
17. Joyner, D. (2015). The Impact of the Students in the Georgia Tech OMSCS [Video]. Presentation given to the GVU Center Brown Bag Seminar series. Retrieved from https://youtu.be/2u2dQOVyen4
18. Kotturi, Y., Kulkarni, C., Bernstein, M. S., & Klemmer, S. (2015). Structure and messaging techniques for online peer learning systems that increase stickiness. In *Proceedings of the Second ACM Conference on Learning @ Scale.* ACM. 31-38.
19. Kulkarni, C., Bernstein, M. S., & Klemmer, S. (2015). PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Proceedings from The Second ACM Conference on Learning @ Scale*. ACM. 75-84.
20. Li, L., Liu, X., & Steckelberg, A. L. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology*, *41*(3), 525-536.
21. Lu, J., & Law, N. (2012). Online peer assessment: effects of cognitive and affective feedback. *Instructional Science*, *40*(2), 257-275.
22. Onah, D. F., Sinclair, J., & Boyatt, R. (2014). Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 Proceedings*, 5825-5834.
23. Paulus, T. M. (1999). The effect of peer and teacher feedback on student writing. *Journal of Second Language Writing*, *8*(3), 265-289.
24. Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013, July). Tuned Models of Peer Assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining*. Memphis, Tennessee.
25. Raman, K., & Joachims, T. (2015). Bayesian Ordinal Peer Grading. In *Proceedings from The Second ACM Conference on Learning @ Scale*. ACM. 149-156.
26. Sankar, P. (2013). Piazza: Our Story [Web Site]. Retrieved from https://piazza.com/about/story
27. Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*(3), 249-276.
28. Tseng, S. C., & Tsai, C. C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education*, *49*(4), 1161-1174.
29. The Sakai Project. (2014). Sakai 10. Retrieved from https://sakaiproject.org/sakai-10
30. Waters, A. E., Tinapple, D., & Baraniuk, R. G. (2015). BayesRank: A Bayesian Approach to Ranked Peer Grading. In *Proceedings of the Second ACM Conference on Learning @ Scale.* ACM. 177-183.
31. Xie, Y., Ke, F., & Sharma, P. (2008). The effect of peer feedback for blogging on college students' reflective learning processes. *The Internet and Higher Education*, *11*(1), 18-25.