# Eroding Investment in Repeated Peer Review: A Reaction to Unrequited Aid?

**David Joyner**
College of Computing
Georgia Institute of Technology
85 5th Street NW, Atlanta, GA 30308
david.joyner@gatech.edu

**Alex Duncan**
College of Computing
Georgia Institute of Technology
85 5th Street NW, Atlanta, GA 30308
alex.duncan@gatech.edu

***Abstract***—Peer evaluation in educational contexts has been well-researched, but there are open questions about the trends students follow over the course of repeated peer evaluation activities within the same course or semester. We hypothesize a trend whereby reviewers who initially heavily invest in giving their classmates strong peer reviews lower their performance over time due in part to disillusionment with the lower-quality feedback they receive. To test this hypothesis, we investigated a dataset of over 50,000 peer assignment evaluations gathered from three semesters of two different courses, totaling over 79 assignments. We examine whether class performance across four quantitative variables drops over the course of the semester, and whether those drops are specifically more prevalent among high-performing reviewers. We find evidence that reviewers who begin the semester committed suffer a greater drop in performance over time, and propose potential causal mechanisms for this drop as well as plans to potentially prevent it.

## INTRODUCTION

Peer evaluation has been well-researched in the learning sciences as a valuable educational exercise both for its instructional value and its secondary benefits to constructing a student community (Boud, Cohen, & Sampson, 2001; Kulkarni, Bernstein, & Klemmer, 2015). The value of the exercise is not only in receiving additional feedback, but also in putting oneself in the position of a critic or teacher and evaluating classmates' work from that perspective (Li, Liu, &

Steckelberg, 2010; Lundstrom & Baker, 2009; Nicol, Thomson, & Breslin, 2014; Rouhi & Azizian, 2013).

A common implementation of this in an authentic classroom setting is to have students perform peer evaluation on multiple assignments throughout the semester; for example, in one of the classes under investigation in this work, students complete 12 assignments, and after each assignment they complete three peer reviews of classmates' submissions. Although peer evaluation as a whole has been well-investigated, this in situ structure is typically not under analysis in existing literature; existing literature focuses more heavily on controlled studies (e.g. Li, Liu, & Steckelberg, 2010; Lundstrom & Baker, 2009). When this structure is analyzed, it is typically taken as a single large-scale intervention (e.g. Kulkarni, Bernstein, & Klemmer, 2015). Topping (1998) provides an overview of the use of peer evaluation in colleges and universities, but very few of the studies referenced examine longitudinal trends with the same students over a semester of repeated peer evaluation; they instead focus on the term-level results of including peer review activities. The question remains largely uninvestigated: how do students' peer evaluation behaviors change within a semester after multiple rounds of giving and receiving evaluations?

Part of this gap is due to the absence of data necessary to investigate this question. Most studies described by Topping (1998) involve evaluations that do not generate accessible ways to analyze reviewer performance; they thus focus more heavily on outcomes as a more accessible data point. However, as more peer evaluation activities have moved online, more data is gathered passively. It is thus now possible to investigate these questions about the changing patterns in student peer evaluations using authentic class data. This is important given the prior observation that the benefit of peer evaluation is more to the reviewer than the recipient (Li, Liu, & Steckelberg, 2010; Rouhi & Azizian, 2013); we ought to be uniquely concerned with reviewers' behaviors as these behaviors are where the desirable outcomes likely arise.

In this work, we investigate these questions with a data set of over 50,000 peer evaluations written in response to essays in two classes in a graduate-level computer science program. Entering this analysis, we hypothesized that we would see two trends: over time, the variation in student behavior will narrow,

and likely decrease overall. To investigate this, we examine multiple objective and quantitative ways of summarizing reviewer behavior in peer evaluation.

## RELATED WORK

This work builds on the existing literature on peer evaluation, but also leverages the theory of performance matching as a potential method for explaining trends in reviewer behaviors.

### Performance Matching

Performance matching is a trend observed in research on group brainstorming and ideation exercises, wherein more engaged and motivated individuals match the performance of their peers over time, leading to an overall decrease in the value of the exercise compared to those individuals acting alone (Paulus & Dzindolet, 1993; Paulus et al., 1996). We hypothesize a similar effect in peer evaluation, whereby individuals more initially invested in reviewing their classmates will lose that investment over time due to a perceived and relative lack of return; their performance will diminish to that of their initially less-invested peers. We also hypothesize a smaller effect whereby initially low-performing students may modestly increase their performance when confronted with the peer reviews generated by their significantly higher-achieving classmates, especially if peer reviews are explicitly graded for substantivity.

### Peer Assessment

Peer evaluation itself has been the subject of an enormous volume of literature documenting its effectiveness both as an educational activity and as a method for generating reliable assessment data. We primarily differentiate two forms of peer assessment: peer feedback or peer review, which focus on the qualitative feedback given, and peer grading, which focuses on the generation of actual values to be used for assigning grades.

First, peer review has been shown to have positive effects on learning outcomes across nearly every topic, context, and medium in which it has been tested: from elementary school (Masten, Morison, & Pellegrini, 1985) to high school (Noonan & Duncan, 2005; Tseng & Tsai, 2007) to college (Dochy, Segers, & Sluijsmans, 1999) and from writing (Wichmann, Funk, & Rummel, 2015) to language learning (Cheng & Warren, 2005) to computer science (Tseng & Tsai, 2007). Multiple

meta-analyses have concluded as such as well (Topping, 1998), and peer review is generally recognized as a highly desirable pedagogical activity.

As the activity has become mainstream, research has been devoted to investigating the mechanics of peer evaluation. Some researchers aim to maximize the benefit associated with peer evaluation, finding that evaluation training can reduce social style bias (May, 2008) and sense-making support can help students apply feedback (Wichmann, Funk, & Rummel, 2015). Others look narrowly at peer evaluation in team projects to identify what expectations students have for evaluations from their teammates (Chen & Lou, 2004; Prins, Sluijsmans, Kirschner, & Strijbos, 2005). With the rise of new technologies, some researchers examine the role of different media or technologies for performing peer evaluation (Cerratto-Pargman, Knutsson, & Karlström, 2015; Pier et al., 2017). This work overlaps with research on evaluations in other domains as well, such as academic peer review and employee performance evaluation. Recent work has investigated the possibility of gender or racial bias in academic peer review, as applied to both authors and reviewers (Bornmann, Mutz, & Daniel, 2007; Borsuk et al., 2009).

Work evaluating the effectiveness in peer grading in assigning scores comparable to expert-generated grades has been more mixed. Several meta-analyses suggest that while it is possible for peer grading to generate marks reliably with expert grading, this only arises under specific conditions relating to both audience and subject matter (Falchikov & Boud, 1989; Falchikov & Goldfinch, 2000). Many recent initiatives in artificial intelligence have aimed to increase the validity of this grade generation (Díez Peláez et al., 2013; Luaces et al., 2015; Staubitz et al., 2016).

Notably, however, no discoverable research has yet looked at longitudinal trends in peer assessment. Most research looks either at a single peer review exercise, such as comparing peer and expert evaluations on a single assignment. The remainder largely looks at the overall effects of peer review over the course of an entire term or course. There are investigations into supporting peer review (Lam, 2010; Liou & Peng, 2009; Min, 2005), including using artificially intelligent support scaffolding (Nguyen, Xiong, & Litman, 2017), but these are generally controlled studies comparing student performance before and after a specific intervention.

### Peer Assessment at Scale

Peer assessment has taken on an added relevance with rise of online education, both in massive open online courses (MOOCs) and in scalable for-credit degree programs. The expectation of essentially infinite scale in MOOCs has forced a significant reliance on peer review for work that cannot be automatically evaluated, and so significant work has been devoted to using peer assessment to support scale (Admiraal, Huisman, & Van de Ven, 2014; Balfour, 2013; Kulkarni, Bernstein, & Klemmer, 2015; Suen, 2014).

On the for-credit side, the emerging field of affordable degrees at scale has used peer review extensively but has cautioned against using peer grading (Joyner, 2017). Instead, this direction generally focuses on the pedagogical value of peer review while reserving formal grade generation for hired graders, although peer review may be used to inform these grades (Joyner et al., 2016).

Although many courses exist in which students complete only one or two peer review assignments, many courses also exist where students are repeatedly engaging in a cycle of peer review. In one study, students completed peer reviews on six assignments during the semester (Joyner et al., 2016). Regardless of whether the purpose of this activity is pedagogical enhancement or grade generation, these activities rely on authentic and earnest engagement by the peer reviewers. Trends in their investment in the activity during the semester may have significant implications for the activity's success for any purpose. Thus, this work builds on existing research by investigating these longitudinal effects: how do students' peer review behaviors change over time? If there are significant trends, then this finding will inform much of the existing literature: algorithms for generating grades must be equipped with the knowledge of the assignment's position in the semester; training for better peer review must target students at the right time before negative habits settle in; and use of repeated peer review must come equipped with incentives to mitigate these trends.

## METHODOLOGY

To understand the dataset and methodology, we will first describe the program from which this data arose. We will then describe the dataset, our hypotheses about it, and our methodology for testing those hypotheses.

**Background**

The evaluations in this study were generated as part of an online Master of Science in Computer Science program. As an online distance learning program, the demographics of the classes are non-traditional: the median age is 35 years old, all students attend school part-time, 85% of students are employed full-time, and the significant majority of the students in the program are domestic students. By contrast, the university's analogous on-campus program has a median age of 23 years old, most students attend school full-time, and the majority are international students (Joyner & Isbell, 2019).

For this study, we selected three full terms from each of two classes in the program that use peer evaluation as part of their standard class administration. These two classes were selected because they attach similar incentive to completing peer evaluation and have maintained relatively stable schedules throughout the three questions under analysis. In each class, students complete an essay assignment almost weekly (between 10 and 16 assignments in the 17-week semester, depending on the course and term). The week following each assignment, students are assigned 3 or 4 of their classmates' assignments to review, which includes grading according to a rubric and providing written feedback. Participating in peer evaluation is a required part of students' participation grades, and students are informed that only substantive evaluations will receive credit. The scores students assign their classmates are not a factor in assigning the evaluation recipient's actual grade; the exercise is solely for the benefit of the participants, not for generating course grades. For this study, a single peer evaluation is one student writing about one classmate's work: thus, each student completes at least 3 to 4 peer evaluations per week. As we correlate performance on these evaluations with in-semester assignment order, this means that each week sees students generating multiple peer evaluations.

Each of the two classes attach an extra incentive to completing peer evaluations early to incentivize rapid feedback: an evaluation completed within 4 days of assignment is worth 1.5 participation points, while a evaluation completed within 7 days is worth 1.0 points. Evaluations completed after 7 days are worth 0.5 participation points. Both classes allow students to give extra peer evaluations above their initially-assigned batch, which are worth 1.0 points each no matter when in the semester they are completed.

## Dataset

The two courses and three terms selected generated a total of 51,966 evaluations across 79 assignments for analysis. Table 1 below shows the breakdown of evaluations per course and per term. Inter-semester differences are due to a combination of differing numbers of assignments and differing enrollment.

**Table 1.** Evaluation counts per term and semester of the two courses under analysis.

| | Term 1 | Term 2 | Term 3 | Total |
|---|---|---|---|---|
| **Course 1** | 8,664 | 9,126 | 9,644 | 27,438 |
| **Course 2** | 8,702 | 7,557 | 8,268 | 24,528 |
| **Total** | 17,366 | 16,685 | 17,915 | 51,961 |

Our independent variable is the order in which an assignment occurred in the semester. For this analysis, we look only at the discrete order and ignore whether off-weeks occurred in between; in doing so, we focus on the effect of repeated exposure to others' evaluations and assignments rather than pure temporal effects. For our dependent variables, we use four quantitative measurements as proxies for commitment:

- Length of the plain text evaluation, referred to as "Length"
- Time spent on the evaluation measured in seconds, referred to as "Duration"
- Meta-evaluation score assigned to the evaluation by its recipient, referred to as "Meta-Score"
- Time before the closest deadline measured in days, referred to as "Earliness"

Among these dependent variables, meta-scores are present for a relatively small fraction of evaluations: only 6,643 of the 51,961 evaluations, 12.8%, received meta-scores. Earliness only applies to evaluations submitted before the early or normal evaluation deadlines and is available for 47,272 evaluations (91.0%). We note that meta-scores are weakly but statistically significantly correlated with evaluation length ($R = 0.2249$, $p < 0.0001$, $n = 6643$, $b = 0.0009$, $a = 5.2317$). Meta-feedback is not significantly correlated with time spent reviewing ($R = 0.0061$, $p = 0.6191$, $n = 6643$) or time before the closest deadline ($R = 0.0247$, $p = 0.0518$, $n = 6199$). Evaluation length and duration are very weakly but significantly correlated ($R = 0.0317$, $p < 0.0001$, $n = 51961$, $b = 0.0027$, $a = 433.9364$),

as are length and earliness ($R = 0.0488$, $p < 0.0001$, $n = 47272$, $b = 16.4531$, $a = 421.7578$). Duration and earliness are not significantly correlated ($R = -0.0094$, $p = 0.0321$, $n = 47272$). All significant correlations are positive, meaning that students who submit earlier and spend longer writing evaluations tend to have longer evaluations, and longer evaluations tend to receive higher meta-scores. It is worth noting that because meta-score is subjective and assigned by the evaluation recipients, it may also be affected by order in the semester: students may get more or less harsh in their judgments as the term progresses.

**Hypotheses & Analysis**

Our primary means of analysis throughout this paper is linear regression: we seek correlations between an assignment's order in the semester and metrics for student commitment on peer evaluations for that assignment. We will look for correlation strength (R, the correlation coefficient) and the slope (b) and intercept (a) of the linear regression equation. Some analyses will base this regression on the entire dataset, while others will focus on aggregated summaries of student activity at particular points in the term.

We entered this analysis with two general hypotheses: first, that commitment to peer evaluation activities would decrease as the semester moves along, and second, that students who began the semester exhibiting high commitment to peer evaluation would diminish more significantly than those who began the semester exhibiting lower commitment. In terms of our dependent variables, we characterize high commitment to peer evaluation as generating longer evaluations, spending more time on evaluations, receiving better ratings from evaluation recipients, and submitting evaluations with more time to spare before the deadline.

To test these hypotheses, we first performed linear regressions between assignment order and all four dependent variables to test for broad correlations. This would test the first hypothesis, that commitment deteriorates class-wide. Second, we examined interquartile ranges within each assignment's evaluations to observe longitudinal trends in the range of the dependent variables. This would test the second hypothesis, that this deteriorating commitment disproportionately took place among the upper quartiles. Third, we labeled students based on the quartile during the first portion of the semester and tracked their performance through the remainder of the semester to observe

within-student trends. This would further test the third hypothesis, that this deterioration was due to diminishing performance from initially high-commitment individuals rather than randomly spread across the class. This third analysis also controls for the possibility that higher-performing students disproportionately skip evaluation tasks later in the semester because their grades are high enough already.

## RESULTS

Below, we will summarize the results of the three analyses separately, and then synthesize these results in the Discussion section of this paper.

### Analysis #1: Overall Trends

To test the first hypothesis regarding overall trends, we performed linear regression analyses between the independent variable (order in the semester) and per-assignment means for each of the four dependent variables. The results of these linear regression analyses are shown below in Table 2.

**Table 2.** Linear regression results between assignment order in the semester and per-assignment means for each dependent variable. Order is placed on the x-axis, while each dependent variable is placed on the y-axis.

| | R | n | b | a |
|---|---|---|---|---|
| Length | -0.7198 | 79 | -18.4464 | 559.9447 |
| Duration | -0.5838 | 79 | -29.7083 | 1058.1470 |
| Meta-Score | -0.3824 | 79 | -0.0424 | 5.8471 |
| Earliness | -0.0751 | 79 | -- | -- |

Notable relationships were observed between three of the four dependent variables and the order of the corresponding assignment in the semester. For example, after spending an average of 1028 seconds on each peer evaluation during assignment 1, students spent on average 30 fewer seconds on the evaluations for each subsequent assignment. Likely the most notable relationship occurs with evaluation length, where the correlation coefficient is the strongest: after writing 541 characters per peer evaluation on the first assignment, students wrote 18 fewer characters for each subsequent assignment's evaluations. In Term 1 of Course 1, for example, this accounted for a drop in average evaluation length from 507 characters on the first assignment to 327 on the last (16th). Due to this stronger relationship with evaluation length and space constraints within this

paper, we will focus our presentation on evaluation length for the rest of the analyses; while length is by no means a perfect proxy for quality, it has been used in the past for useful results (Kulkarni, Bernstein, & Klemmer, 2015).

Notably, while length and order are correlated for both courses, the relationship is stronger and more negative in Course 2 than 1. Table 3 separately calculates linear regression formulae for the two courses.

**Table 3.** Linear regression results between assignment order in the semester and per-assignment mean evaluation length, separated by course..

| | R | n | b | a |
|---|---|---|---|---|
| Course 1, Length | -0.6417 | 35 | -9.7164 | 458.8929 |
| Course 2, Length | -0.8772 | 44 | -26.3486 | 655.9640 |

We hypothesize that this difference is due to differences in each class's assignment directions. Course 1 has more freedom for individual variation in the assignments, and therefore we hypothesize some students may be more motivated to continue participating earnestly in peer evaluation as each assignment is more unique. More importantly, however, this asserts that the trends are present separately for two different courses, although their specific trajectories may differ.

**Analysis #2: Interquartile Trends**

Analysis #1 confirms the hypothesis that commitment to peer evaluation (at least as measured by our dependent variables) declines as the semester progresses. However, this could happen for multiple reasons: all students may drop off evenly, low-performers may quit altogether, or high-performers may become disproportionately more disillusioned by a perceived one-sidedness of their participation. We hypothesized this third trend: that those who started the semester most committed to peer evaluation diminished more significantly, due at least in part to the perceived lack of reciprocation.

To test this hypothesis, for each assignment we divided the observed peer reviews for that assignment into four groups based on length: the top (fourth) quartile were the longest 25% of the reviews on that assignment, while the bottom (first) quartile held the shortest 25%. We selected four groups due to the diminishing differences between the groups as the number of potential groups expanded; while the third and fourth quartiles have clear differences, expanding

to five or six groups created lower groups with miniscule differences in measured performance.

Equipped with these quartiles, we again performed linear regressions on the divisions between quartiles to see if more significant changes were observed in some quartiles than others. If the first hypothesis above was true, we would expect all to drop relatively evenly; if the second hypothesis was true, we would expect the bottom quartile to drop more notably than the top; and if the third hypothesis was true, we would expect the top quartile to drop more notably than the bottom. The results of this analysis are shown in Table 4.

**Table 4.** Linear regression ranges between assignment order in the semester and per-assignment length quartiles, both in aggregate and separated by course.

| Course | Value | R | n | b | a |
|---|---|---|---|---|---|
| Both Courses | 1st/2nd Quartile Border | -0.8315 | 79 | -10.8631 | 286.8559 |
| | 2nd/3rd Quartile Border | -0.7675 | 79 | -17.3460 | 476.8819 |
| | 3rd/4th Quartile Border | -0.7138 | 79 | -26.2704 | 738.3144 |
| Course 1 | 1st/2nd Quartile Border | -0.7978 | 35 | -8.2138 | 260.2332 |
| | 2nd/3rd Quartile Border | -0.7071 | 35 | -10.4894 | 400.0160 |
| | 3rd/4th Quartile Border | -0.6331 | 35 | -14.2049 | 593.8034 |
| Course 2 | 1st/2nd Quartile Border | -0.8800 | 44 | -12.9226 | 310.5665 |
| | 2nd/3rd Quartile Border | -0.8887 | 44 | -23.1255 | 548.1575 |
| | 3rd/4th Quartile Border | -0.8812 | 44 | -36.9626 | 875.2479 |

The results of this analysis support our hypothesis. The borders between each pair of quartiles decline on each subsequent assignment, which means that the decrease in the average evaluation length over the course of the semester arises from across the spectrum. More importantly, however, the slopes associated with the higher quartiles are more negative than the slopes associated with the lower quartiles. For all courses, the border between the first and second quartiles decreases by 10 characters per assignment, while the border between the third and fourth quartiles decreases by 26 characters per assignment. As a result, the interquartile range shrinks from assignment to assignment as well: the median length of the longer evaluations moves 15 characters closer to the median length of the shorter evaluations per assignment.

As before, this trend exists for both courses, but to different extents. Course 2 shows a much greater correlation between assignment order and each of the three quartile boundaries, as well as greater slopes: the interquartile range shrinks by 24 characters per assignment for Course 2, compared to 6 characters per assignment for Course 1. Although Course 2 has longer peer evaluations overall than Course 1, this greater decline is still relatively stronger: the interquartile range shrinks by 4% of the initial average evaluation length in Course 2, compared to 2% of the initial average evaluation length in Course 1.

Chart 1 shows these trends in a visual manner. The larger chart on the left aggregates all courses and terms, while the individual charts to the right show the trends specific to each course and term. We performed the same analysis on Duration and found similar correlations. The border between the third and fourth quartiles dropped on average 20 seconds per assignment, while the border between the first and second quartiles dropped on average 10 seconds per assignment. We also performed the same analysis on Meta-Scores, but did not find statistically significant correlations.

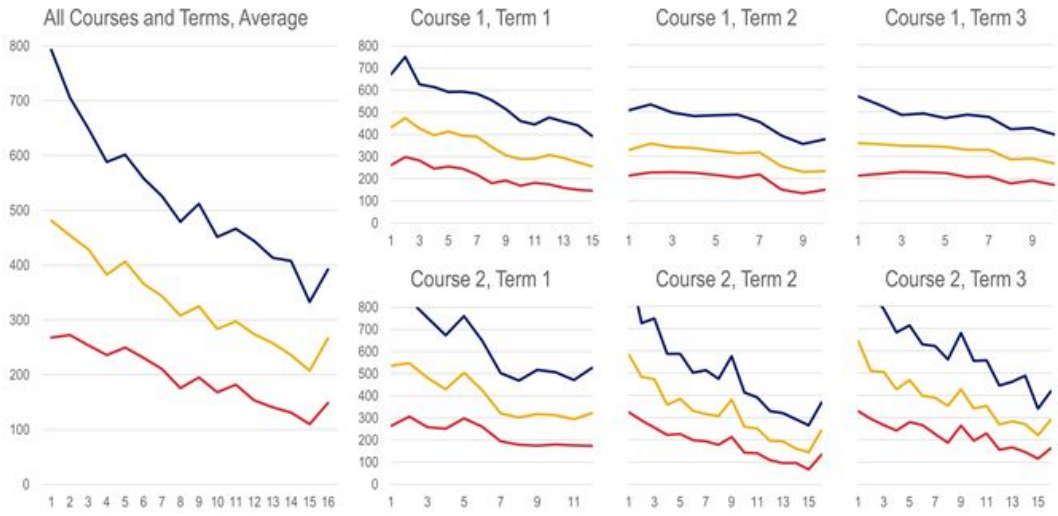


**Chart 1.** Interquartile boundaries for evaluation lengths for each course and term, as well as all courses and terms averaged together. Red (bottom) lines separate the first two quartiles; yellow (middle) lines separate the second and third quartiles; blue (top) lines separate the third and fourth quartiles. The vertical axis is average evaluation length, and the horizontal axis is assignment order in the term.

### Analysis #3: Individual Trends

Analysis #2 supports the hypothesis that the interquartile range of evaluation lengths diminishes over time due to the median length of longer evaluations

falling more than the median length of shorter evaluations. It is not inherently clear, however, if these trends are specific to individuals or products of the aggregated class. For these results to indicate specifically these decreases were borne by initially high-achievers, we would expect to see students who initially demonstrate high commitment to themselves show a greater decrease in performance.

For this final analysis, then, we followed a similar pattern to Analysis #2. However, instead of looking at the interquartile ranges for the classes' performance as a whole, we instead followed the trajectories of individual students. We first calculated an initial quartile for each student based on their evaluations on the first two assignments. We then traced the mean evaluation length for students within each quartile over the course of the semester. In this way, we can more specifically see the trajectories of those individual students who demonstrated high and low commitment to start the semester. Because students can complete as few as three evaluations per assignment, we chunked assignments into pairs to better account for occasional outliers from individuals (such as an individual submitting a short evaluation because of an error reading the paper). The results of this analysis are shown in Table 5.

These results support the hypothesized changes in performance over time. The rank-order of student performance remained consistent throughout the semester: for nearly every pair of assignments and quartiles (117 of 120), the average length of evaluations written by students initially in the fourth (top) quartile was above that of students initially in the third (second-highest) quartile, which was above that of students initially in the second (second-lowest) quartile, which was above that of students initially in the first (bottom) quartile. Despite this, students in the top quartile on the first pair of assignments saw a steeper decline. In Course 2, students in the top quartile for length on the first two assignments wrote on average 110 fewer characters per evaluation on each subsequent pair of assignments, while students initially in the bottom quartile wrote only 10 fewer characters on each subsequent pair. Chart 2 depicts these trends visually.

**Table 5.** Linear regression results for Analysis #3. Here, the independent variable (x-axis) is assignment order in the semester. The dependent variable (y-axis) is mean evaluation length, separated out based on student performance on the first two assignments of the semester. First quartile students, for each, are those whose mean evaluation lengths were in the bottom 25% of the class on the first two assignments.

| Course | Value | R | n | b | a |
|---|---|---|---|---|---|
| Both Courses | First (Bottom) Quartile | -0.4316 | 40 | -8.3018 | 238.5612 |
| | Second Quartile | -0.7180 | 40 | -21.2425 | 397.6283 |
| | Third Quartile | -0.7476 | 40 | -40.3907 | 589.7202 |
| | Fourth (Top) Quartile | -0.7129 | 40 | -79.1557 | 932.0373 |
| Course 1 | First (Bottom) Quartile | -0.5198 | 18 | -6.9610 | 223.7788 |
| | Second Quartile | -0.6559 | 18 | -14.3000 | 344.6272 |
| | Third Quartile | -0.6201 | 18 | -20.4708 | 484.8481 |
| | Fourth (Top) Quartile | -0.6916 | 18 | -45.2756 | 720.9652 |
| Course 2 | First (Bottom) Quartile | -0.4569 | 22 | -10.1541 | 254.4639 |
| | Second Quartile | -0.8879 | 22 | -28.5899 | 451.2246 |
| | Third Quartile | -0.8862 | 22 | -56.9730 | 685.8623 |
| | Fourth (Top) Quartile | -0.8485 | 22 | -110.476 | 1135.493 |

These visual depictions clarify the trend observed. On average, students in the top quartile by length on the first pair of peer evaluations wrote 757 more characters than students in the bottom quartile; by the second pair, that dropped to 511 more characters, and by the final pair, it had dropped to 198 more characters.

We also performed this analysis on Duration, assigning students to an initial quartile based on the time spent on their evaluations for the first pair of assignments. Duration saw similar correlations with students initially in the third and fourth quartile, but not those initially in the first and second quartile. Students in the third quartile spent an average of 50 fewer seconds per evaluation on each subsequent pair of assignments, while students in the fourth quartile spent 80 fewer seconds. Not enough meta-scores were given out on the first pair of assignments to categorize students into initial quartiles and perform this analysis.

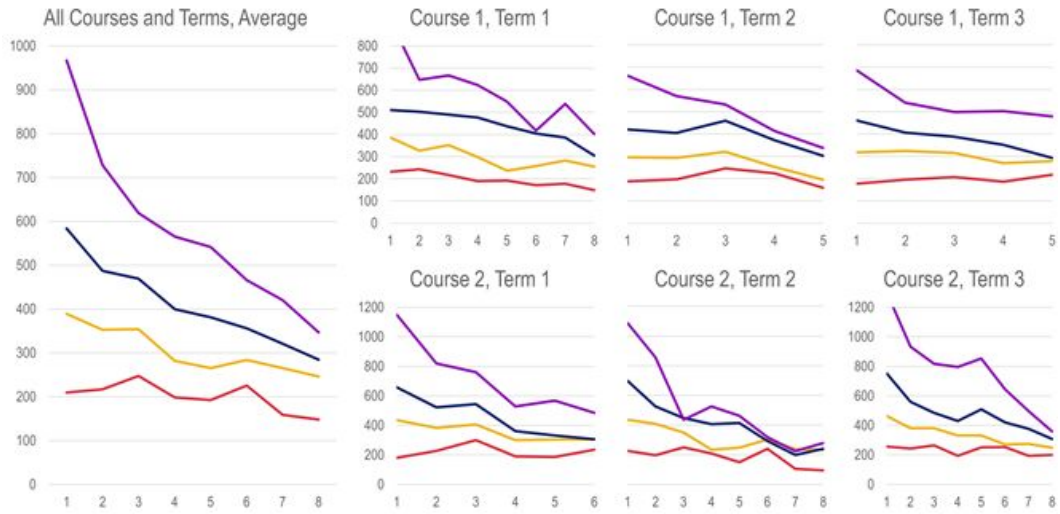


**Chart 2.** Mean evaluation lengths per pair of assignments by students based on their quartile by length on the first pair of assignments. Purple (top) lines are students in the top quartile on the first two assignments; blue lines, the third quartile; yellow lines, the second quartile; and red lines are students in the bottom quartile.

### Analysis Discussion

The above three analyses identified three related trends in the dataset. First, as the semester moves forward and students complete peer evaluations, their overall performance drops: they write shorter evaluations, spend less time on their evaluations, and their classmates assign lower meta-scores to the evaluations that they receive. Second, this trend is borne disproportionately: the average evaluation length and the average time spent on peer evaluations drops more significantly for the upper quartiles than the lower quartiles. Phrased more simply, long evaluations get shorter faster than short evaluations. This trend is true both in absolute terms and as a percentage of prior review length, indicating it is not an instance of all reviews dropping by the same relative length. This trend leads to a much smaller interquartile range for both evaluation length and time spent per evaluation on later assignments than earlier assignments. Finally, it is also more specifically true that it is the individuals who write the longest evaluations and spend the most time early in the term that see their average length drop the most over the course of the semester. It is notable that the trend was present in each of two classes and each of six class-semesters, but to varying degrees, indicating that there are inter-class differences to these trends, but that the trends are not specific to only one class.

**Hypothesized Cause**

Our hypothesized cause of this trend is something resembling performance matching: initially high-achieving participants lower their commitment to peer review over time due to the perception that the aid they gave their classmates was unrequited; those students writing the longest reviews are statistically guaranteed to mostly receive reviews that are not as thorough as the ones they give. To further prove this hypothesis, we would need to statistically correlate the length of reviews received with the length of reviews given over time, especially to identify if there are students whose performance increases (contrary to the overall trends) because, due to random chance, they received on average largely longer reviews than they give initially. This, along with more rigorous mechanisms for evaluating review quality, is planned for future work.

There are other hypotheses for these trends as well. One hypothesis is that performance naturally diminishes over time, and this would occur regardless of the quality of feedback students receive. However, this hypothesis does not explain why initially high-performing reviewers drop more precipitously as a percentage of their initial performance. Another hypothesis is that over time, higher-achieving students become more aware of how little they can get away with and still receive credit; however, students do not learn until 5-6 weeks after a peer review has been given if they have received credit, and so this hypothesis would not explain the initial steep drop. Similarly, students do not learn if their classmates' reviews received credit, and so high-achieving students cannot assume that the poor feedback they receive is receiving credit. Finally, it may be hypothesized that a drop in review length and time spent giving feedback are due to improvements in giving good feedback efficiently, or in improvements to the assignments receiving feedback; this, however, does not explain the relative drop in meta-score.

For these reasons, we find the most compelling explanation for the observed trend to be a reaction by high-performing students to unrequited effort in the peer reviews they receive compared to the ones they give.

**Possible Solutions**

Regardless of the cause, this analysis demonstrates a trend educators would likely want to resolve: if we can encourage high-performing peer reviewers to persist in giving high-quality peer reviews, we may improve the learning

experience for everyone, and at scale. The same may be said for encouraging low-performing peer reviews to invest more highly, but that goal is more persistent in the existing literature; we know of no present work on preserving the commitment of pre-existing high-performing peer reviewers.

Toward this end, we propose three possible solutions to this trend. The first entails more intelligent matching. Probability dictates that many peer reviewers in the top quartile will receive strictly inferior peer reviews. Intelligent and dynamic matching can ensure that high-performers are consistently partnered with at least one other high-performer, providing greater requitement of their efforts.

A second, more extreme solution may be to *exclusively* partner peer review performers with similar classmates. This turns peer review into a more straightforward investment: the better feedback you provide, the better feedback you will receive in a more directed and deliberate manner. However, this comes with a significant cost: one major benefit of peer review is that it partners novices and experts together. The benefit is mutual, but likely borne more by the novices. Partnering exclusively within performance groups loses this potential benefit.

A third solution may be to give greater differential credit based on review quality. In this study, reviews either receive credit or they do not, and the range between the best peer reviews and the worst credit-receiving reviews is large. In the studies above, over 99.9% of reviews in the top three quartiles by length receive credit, and over 75% of the reviews in the bottom quartile also receive credit. However, if differential credit were given such that a great review receives more credit than a good review, high-performing reviewers may perceive the score as sufficient compensation to persist in investing in peer review. This, however, presents other issues, such as justifying to students why points were "deducted" for good-but-not-great peer reviews, and scaling assessment to grade tens of thousands of peer reviews. These procedural challenges likely prevent this solution from being used in the sort of at-scale environment in which this study was conducted.

## CONCLUSION

This analysis has demonstrated that in two separate courses with repeated peer evaluation activities, those students who initially invest heavily in giving good peer reviews disengage over time. This disengagement is more significant than

that experienced by their initially less-motivated classmates. As a result, the range of observed values at the end of the semester is significantly narrower than at the beginning.

This finding has multiple implications. First, specific to courses with repeated peer evaluation, steps ought to be taken to limit this performance matching and maintain the initial motivation of more highly-engaged students. Otherwise, peer evaluation risks becoming ineffective by the end of the term. Second, for courses with other socially visible behaviors, similar performance matching may also occur. There are many other mechanics whereby students are aware of their classmates' behaviors, such as forum-posting and class attendance. It is worth checking for these trends in any such activity where students may be peripherally aware of their classmates' performance.

**Limitations**

There are a number of limitations to the generalizability of this study. First and foremost, the student body under analysis here is non-traditional: the students have a median age of around 35, most are working full-time, and they generally have richer professional backgrounds. We speculate that this might lead to the early-term separation being more pronounced than it might be in more traditional classes, especially as the top quartile behaves more disproportionately well. Second, while these classes are both essay-based, they are nonetheless in a computer science graduate program; neither the assignments themselves nor the exercise of peer evaluation of free writing are natural fits for this audience, and student bodies more accustomed to this type of work may behave differently. Finally, these are entirely online classes: peer evaluation occurs asynchronously in part due to the asynchronous nature of online work. These trends may differ for in-person classes where social relationships and pressures to perform beneficially are stronger, or for synchronous peer review exercises where the audience is more captive and where there are more empathy-driven cues to participate authentically.

In addition to the limitations to the generalizability of these conclusions, there are also potential alternate explanations for these trends that are worth exploring. For example, it may be plausible that early long reviews were driven more by need than by reviewer commitment; if students learned to write better assignments from the feedback they received during the first few, it would be

reasonable that shorter corrective reviews would arise. We do not hypothesize that this is the case as peer review in this context is deliberately more constructive than corrective; these are exploratory essays, and peer reviews are expected to contribute to a discussion rather than pass judgment. Nonetheless, future work may evaluate the types of feedback present in these peer reviews to gauge whether positive and collaborative feedback is remaining high while corrective feedback diminishes, suggesting learning rather than performance matching.

**Future Work**

Aside from addressing these limitations and testing the generalizability of these findings, future work in this area aims to better-quantify these trends and experimentally test approaches for mitigating them. First, while this work has used evaluation length and time spent writing an evaluation as proxies for quality, these are not perfect metrics; it would be preferable to have expert-assigned ratings as outcome variables, although as noted manual grading of these reviews is intractable. Since this study was conducted, a machine learning classifier has been deployed to the peer review system which assigns automated ratings to peer reviews; this outcome variable will be used in future analyses as well.

Secondly, the next question from this analysis is: how can we mitigate these trends? Can we limit the extent to which performance matching occurs? Can we incentivize initially lower-performing students to raise their performance so that performance meets in the middle rather than approaching the lower quartile? Fortunately, the repeated structure of the classes under analysis in this study means that future semesters may be directly comparable to assess our progress in combating these trends. In future terms, we plan to experiment with the solutions posed above regarding more intentional assignment of reviewers to reviewees based on past peer review performance. We hope this introduces greater rewards for higher-performing reviewers and encourages them to maintain their high performance. We have also considered merging peer and grader evaluation workflows in hopes that perceiving grader and peer evaluation in the same interface may incentivize greater commitment to peer evaluation. Finally, we also plan to share exemplary peer evaluations in addition

to exemplary assignments to introduce an additional recognition for initially highly-committed students.

**ACKNOWLEDGEMENTS**

We are grateful to the architects of the peer feedback tool used in the program for their consistent commitment to supporting research in peer review through numerous custom enhancements and features. We are also grateful to the staff and administration behind this program for supporting research into its inner workings.

**REFERENCES**

1. Admiraal, W., Huisman, B., & Van de Ven, M. (2014). Self-and peer assessment in massive open online courses. *International Journal of Higher Education, 3*(3), 119-128.
2. Balfour, S. P. (2013). Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™. *Research & Practice in Assessment, 8*, 40-48.
3. Bornmann, L., Mutz, R., & Daniel, H. D. (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics, 1*(3), 226-238.
4. Borsuk, R. M., Aarssen, L. W., Budden, A. E., Koricheva, J., Leimu, R., Tregenza, T., & Lortie, C. J. (2009). To name or not to name: The effect of changing author gender on peer review. *BioScience, 59*(11), 985-989.
5. Boud, D., Cohen, R., & Sampson, J. (2001). *Peer Learning in Higher Education: Learning from & with Each Other*. Psychology Press.
6. Cerratto-Pargman, T., Knutsson, O., & Karlström, P. (2015). Materiality of Online Students' Peer-Review Activities in Higher Education. In *11th International Conference on Computer Supported Collaborative Learning* (pp. 308-315). International Society of the Learning Sciences.
7. Chen, Y., & Lou, H. (2004). Students' perceptions of peer evaluation: An expectancy perspective. *Journal of Education for Business, 79*(5), 275-282.
8. Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing, 22*(1), 93-121.
9. Díez Peláez, J., Luaces Rodríguez, Ó., Alonso Betanzos, A., Troncoso, A., & Bahamonde Rionda, A. (2013). Peer assessment in MOOCs using preference learning via matrix factorization. In *NIPS Workshop on Data Driven Education*.

10. Dochy, F. J. R. C., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24*(3), 331-350.
11. Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395-430.
12. 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.
13. Joyner, D. A., Ashby, W., Irish, L., Lam, Y., Langston, J., Lupiani, I., Lustig, M., Pettoruto, P., Sheahen, D., Smiley, A., Bruckman, A., & Goel, A. (2016, April). Graders as Meta-Reviewers: Simultaneously Scaling and Improving Expert Evaluation for Large Online Classrooms. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 399-408). ACM.
14. Joyner, D. A. (2017, April). Scaling Expert Feedback: Two Case Studies. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale* (pp. 71-80). ACM.
15. Joyner, D. A. & Isbell, C. (2019). Master's at Scale: Five Years in a Scalable Online Graduate Degree. In *Proceedings of the Sixth Annual ACM Conference on Learning at Scale*. Chicago, Illinois, USA.
16. Kulkarni, C., Bernstein, M. S., & Klemmer, S. (2015). PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Proc. from The Second ACM Conference on Learning @ Scale*. ACM. 75-84.
17. Lam, R. (2010). A peer review training workshop: Coaching students to give and evaluate peer feedback. *TESL Canada Journal, 27*(2), 114.
18. 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.
19. Liou, H. C., & Peng, Z. Y. (2009). Training effects on computer-mediated peer review. *System, 37*(3), 514-525.
20. Luaces, O., Díez, J., Alonso-Betanzos, A., Troncoso, A., & Bahamonde, A. (2015). A factorization approach to evaluate open-response assignments in MOOCs using preference learning on peer assessments. *Knowledge-Based Systems, 85*, 322-328.
21. Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language writing, 18*(1), 30-43.

22. Masten, A. S., Morison, P., & Pellegrini, D. S. (1985). A revised class play method of peer assessment. *Developmental Psychology, 21*(3), 523.
23. May, G. L. (2008). The effect of rater training on reducing social style bias in peer evaluation. *Business Communication Quarterly, 71*(3), 297-313.
24. Min, H. T. (2005). Training students to become successful peer reviewers. *System, 33*(2), 293-308.
25. Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education, 39*(1), 102-122.
26. Noonan, B., & Duncan, C. R. (2005). Peer and self-assessment in high schools. *Practical assessment, research and evaluation, 10*(17), 1-8.
27. Nguyen, H., Xiong, W., & Litman, D. (2017). Iterative design and classroom evaluation of automated formative feedback for improving peer feedback localization. *International Journal of Artificial Intelligence in Education, 27*(3), 582-622.
28. Paulus, P. B., & Dzindolet, M. T. (1993). Social influence processes in group brainstorming. *Journal of Personality and Social Psychology, 64*(4), 575.
29. Paulus, P. B., Larey, T. S., Putman, V. L., Leggett, K. L., & Roland, E. J. (1996). Social influence processing in computer brainstorming. *Basic and Applied Social Psychology, 18*(1), 3-14.
30. Pier, E. L., Raclaw, J., Ford, C. E., Kaatz, A., Carnes, M., & Nathan, M. J. (2017). Videoconferencing in Peer Review: Exploring Differences in Efficiency and Outcomes. In *12th International Conference on Computer Supported Collaborative Learning. International Society of the Learning Sciences*.
31. Prins, F. J., Sluijsmans, D. M., Kirschner, P. A., & Strijbos, J. W. (2005). Formative peer assessment in a CSCL environment: A case study. *Assessment & Evaluation in Higher Education, 30*(4), 417-444.
32. Rouhi, A., & Azizian, E. (2013). Peer review: Is giving corrective feedback better than receiving it in L2 writing?. *Procedia-Social and Behavioral Sciences, 93*, 1349-1354.
33. Staubitz, T., Petrick, D., Bauer, M., Renz, J., & Meinel, C. (2016, April). Improving the peer assessment experience on MOOC platforms. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale* (pp. 389-398). ACM.

34. Suen, H. K. (2014). Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning, 15*(3).
35. Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276.
36. 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.
37. Wichmann, A., Funk, A. L., & Rummel, N. (2015). Maximizing benefit of peer-feedback to increase feedback uptake in academic writing. In *11th International Conference on Computer Supported Collaborative Learning*. International Society of the Learning Sciences.