# Grading the Graders: Motivating Peer Graders in a MOOC

Yanxin Lu, Joe Warren, Chris Jermaine, Swarat Chaudhuri and Scott Rixner
Department of Computer Science
Rice University
Houston, TX, 77007
{yanxin.lu,jwarren,cmj4,swarat,rixner}@rice.edu

## ABSTRACT

In this paper, we detail our efforts at creating and running a controlled study designed to examine how students in a MOOC might be motivated to do a better job during peer grading. This study involves more than one thousand students of a popular MOOC. We ask two specific questions: (1) When a student knows that his or her own peer grading efforts are being examined by peers, does this knowledge alone tend to motivate the student to do a better job when grading assignments? And (2) when a student not only knows that his or her own peer grading efforts are being examined by peers, but he or she is also given a number of other peer grading efforts to evaluate (so the peer graders see how other peer graders evaluate assignments), do both of these together tend to motivate the student to do a better job when grading assignments? We find strong statistical evidence that "grading the graders" does in fact tend to increase the quality of peer grading.

## 1. INTRODUCTION

A *massive open online course* (MOOC) [24] is a web-based online course that is open to virtually all interested participants, with few restrictions to registration. Running a MOOC is challenging for many reasons, not the least of which is ensuring high-quality grading. The largest MOOCs have on the order of tens of thousands of participants who actually complete some fraction of the assignments and examinations. Clearly, grading at this scale is beyond even the largest team of instructors and TAs.

One solution to the problem of grading so many submissions is for the instructor to design assignments and examinations in such a way as to ensure that student work can be automatically graded. For example, multiple choice examinations can be given. For another example, in a computer programming class, student programs can be put through a test suite by an automated grader, and the number of test cases passed can be used to assign a student a grade.

The obvious problem with automated grading is that there exist courses for which purely automated grading is not possible. An example that we are intimately familiar with (and the course that is the subject of the experimental study described in this paper) is a course on interactive game programming in Python[1]. Since games by their very nature require a user to play them in order to test correctness, it is exceedingly difficult to utilize automated grading to score student-constructed games. Another example is a mathematics class where the construction of proofs is a key part of the course; fully automated grading of student proofs is not feasible at this time. Courses in the social sciences or languages are also not amenable to fully automated grading.

**Peer Grading.** The common way to scale up grading that cannot be automated is to rely on *peer grading* [21, 9, 30, 22, 8, 14, 15], where student submissions are distributed to other students in the class to be graded. Peer grading has many benefits. In addition to offering a way to crowdsource the grading of student work so as to achieve virtually infinite scalability, peer grading also has pedagogical benefits [6]. Students who grade other students' assignments can benefit from this activity. In particular, they are forced to carefully consider the validity of a wide variety of solutions to a problem. As a result, commonly used MOOC platforms such as Coursera [1] offer built-in facilities to support peer grading.

Peer grading, however, is not a panacea. Our personal experience is that in a MOOC, low quality peer grading can be a problem. The main reason for poor grading seems to be a simple lack of effort, rather than inability or maliciousness. On Coursera, for example, students are assessed a 20% penalty when they fail to peer grade. One unfortunate (and common) student response is simply to give all peers perfect scores. Consider the "An Introduction to Interactive Programming in Python" (IIPP) course that we have offered for a number of years on Coursera, and which serves as the subject of the study described in this paper. On the two assignments central to our study (Stopwatch and Memory), in those cases where we were able to automatically find an error that should have resulted in one or more points being deducted, the peer grader gave full credit 53% of the time (4,168/7,855 grades given). In contrast, in those cases where we could find no error automatically, only 2% of the time (292/16,427 grades given) did the grader take off more than one point.

This sets up the question that is at the heart of the paper:

*How can an educator running a MOOC motivate students to do a better job of peer grading?*

Our goal is to identify a simple and practical method for motivating students to perform high-quality peer grading, and to then rigorously test this method in a controlled experiment, in a real MOOC.

While we are primarily interested in MOOCs, we point out that answering this question definitively would have implications beyond MOOCs. Consider the task of ensuring high-quality peer re-

[1]See https://www.coursera.org/course/interactivepython.


*WWW 2015,* May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741649.

views of submissions to a competetive, academic conference. Reviewers agree to review submissions out of a sense of obligation or because they want to be associated with a prestigious conference, and often lack motivation to do a good job. Methods that work well in a MOOC might also help in motivating conference reviewers.

**Motivating Peer Graders.** One method for motivating peer graders that seems to have entered into the folklore is the method of "sentinels." Instructors or TAs pre-grade a few assignments, which are then added into the general population of assignments. These are called "sentinals." Since we know something of the ground-truth grade for the sentinels, it is possible to identify graders whose grades differ significantly from the expected score. Such off-target grading efforts could be identified, and students might be given some sanction for doing such a poor job grading the sentinels.

However, there are some inherent problems with the use of sentinels. Based on our experience in delivering several MOOCs, a substantial subset of the students enjoy discussing their peers' work in the class forums. Eventually, the sentinels themselves would be a popular topic for discussion. Of course, this knowledge would diminish the effectiveness of sentinels in motivating student effort during peer grading. Further, sentinels are meant to detect and presumably punish poor grading efforts. The idea of punishing students for poor grading seems counter-productive, especially when many students who are taking the MOOC are doing so simply to learn the topic. How does one punish a student who simply wants to learn the course material?

Perhaps a better approach would be to motivate students to do a better job. In this paper, we investigate a very different method for motivating peer graders. Rather than punishing (or rewarding) students for their peer grading, we examine the utility of crowdsourced examination of peer grading efforts. We ask two questions:

1. When a student knows that his or her own peer grading efforts are being examined by peers, *does this knowledge alone* tend to motivate the student to do a better job when grading assignments?

2. When a student not only knows that his or her own peer grading efforts are being examined by his or her peers, but he or she is also given a number of other peer grading efforts to evaluate (so the students see how other peer graders evaluate assignments), *do both of these together* tend to motivate the student to do a better job when grading assignments?

Crucially, there is no punishment or reward under either regime. There is only the knowledge that one's peers are going to examine one's grading efforts (case 1), and in addition, an exposure to how other peer graders have evaluated assignments (case 2).

**Our Contributions.** The idea of "grading the graders" is not revolutionary. In the context of peer review of scientific papers, this is often mentioned as a possible mechanism for ensuring high-quality peer review. However, most or all of the ideas along these lines rely on logical argument or thought experiment to demonstrate utility. Our efforts differ in that not only have we designed and implemented a system for grading the graders, but—far more importantly—we have also designed and run a large-scale, controlled experiment to evaluate the utility of the approach. Such studies should be considered mandatory as the community attempts to figure out the correct way to run a MOOC.

Our specific contributions are as follows:

1. We describe two easily-implementable and easily-deployable methods for motivating peer graders in a MOOC. These methods have the advantage that they require only moderate levels of effort from the MOOC community.

2. We conduct a carefully designed, controlled experimental evaluation of these methods in the context of a popular interactive (game programming) MOOC. We assert that such a controlled experiment is the only reliable way to collect data supporting or refuting the utility of a particular methodology in the context of a MOOC.

3. We find evidence that there are significant differences among the various study groups, even among the subset of students who are motivated enough to sign up for such a study.

**Summary of Findings and Recommendations.** Surprisingly, we found little evidence that simply knowing that one's grading efforts were going to be graded results in a superior grading effort. That is, those study participants who did not grade others, but only had their grading efforts graded, did not perform much better than those in the control group.[2] However, those who participated in the full regime—students who had their grading efforts graded *and* graded others' grading efforts—not only did a better job grading during the study (which lasted for two assignments), *but the positive effects were lasting*. That is, those who participated in the "grading the graders" regime continued to do a better job than those who did not participate in the full regime, even after the study ended.

Thus, the key to achieving better grading results seems to be actually seeing how other people grade, and not simply knowing that grading efforts are being monitored. As discussed in Section 6, we conjecture that actually seeing that other students put in the effort to do a good job helps provide the motivation that students need to do a good job when it is their turn to grade. In this case, simply showing examples of good grading is not enough; *students must be shown evidence that their peers are actually producing such high quality efforts*. This is exactly what the "grading the graders" regime does. Thus, our recommendation is that MOOCs that rely on peer grading should utilize the "grading the graders" regime for at least one or two assignments at the beginning of the class; our study seems to indicate that this should have a positive effect on grading quality.

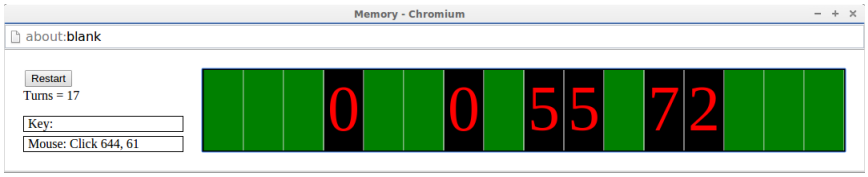
## 2. BACKGROUND

IIPP is a 9-week course designed to help students with very little computing background learn the basics of writing simple interactive programs (games of various types) using the Python programming language. No prior programming experience is assumed. The course covers the basic syntax and semantics of Python, as well as object-oriented programming. Students learn the material of the course by completing a series of "mini-projects" including games such as Pong, Blackjack and Asteroids. We administered the study in the Fall 2013 session of the course. In that session, 120,000 students enrolled and 7,500 finished the course.

One challenging aspect of designing a Python programming MOOC for beginners is making sure that tens of thousands of inexperienced students are able to develop and run Python programs on their own machines. These students use a wide variety of machines and operating systems and there is no way of ensuring that they will all be willing or able to install the same version of any particular software package. Since IIPP relies on peer grading, these programs need to be completely portable. To facilitate this, IIPP

[2]As an aside, this would seem to argue *against* a sentinel-based approach, which by its nature relies on monitoring of grading efforts to increase grading quality.


(a) Screen shot of Stopwatch assignment solution.




(b) Screen shot of Memory assignment solution.

Figure 1: Two IIPP programming assignments that are the subject of the study.

makes use of a browser-based programming environment called "CodeSkulptor" [29]. Any instructor or peer grader with access to the URL that points to the saved code can open and run the code, meaning that submitting a program is equivalent to submitting a URL.

```
import simplegui
# define global variables
minute = 0
second = 0
millisecond = 0
time = "0:00.0"
score = "0/0"
wins = 0
attempts = 0

#Function for the game
def score_keeper():
    ...

# define event handlers for buttons; "Start", "Stop", "Reset"
def timer_handler():
    ...

#Ensure there is a zero before the one
def second_string():
   ...

def reset_handler():
    timer.stop()
    time = "0:00:0"
    second = 0
    millisecond = 0
    minute = 0
    score = "0/0"
    attempts = 0
    wins = 0

# define draw handle
def draw_handler(canvas):
     global time
     canvas.draw_text(time, (100, 150), 50, 'White')
     canvas.draw_text(score, (230, 30), 30, 'Green')

# register event handlers
def start_handler():
     timer.start()

def stop_handler():
     timer.stop()
     score_keeper()

# create frame
frame = simplegui.create_frame("Stopwatch THE GAME", 300, 300)
button1 = frame.add_button('Start', start_handler, 100)
button2 = frame.add_button('Stop', stop_handler, 100)
button3 = frame.add_button('reset', reset_handler, 100)
timer = simplegui.create_timer(100, timer_handler)
frame.set_draw_handler(draw_handler)
# start frame
frame.start()
```

Figure 2: Source code listing of a submitted Stopwatch submission.

## 2.1 Assignments

The study described here concerns two different IIPP programming assignments, Stopwatch and Memory. We describe those programming assignments now.

**Stopwatch.** Stopwatch was assigned during the fourth week of the course. 10,500 students completed the regular version of the Stopwatch assignment, and 2,366 additional students participated in the study and completed the study version of the assignment.

In this project, students write the application whose screenshot is shown above in Figure 1a. Students must implement three buttons: a "start" button, a "stop" button, and a "reset" button. The application implements a simple game where a player presses the start button, which starts a timer that is accurate down to a tenth of a second. The player then attempts to press the stop button when the tenths position of the timer is zero (that is, the player attempts to hit "stop" at a whole second). The application should display the number of times that the user does this correctly. For example, displaying "3/4" means that the user has successfully stopped the timer at a whole second three out of four times. A partial listing of one student's code for this assignment is given in Figure 2.

**Memory**. Memory was assigned at the beginning of the sixth week of IIPP. 7,600 students completed the regular version of Memory, and 1,746 of the 2,366 students who completed the study version of Stopwatch also completed the study version of memory.

In this project, students build a game which first displays eight pairs of cards face down. A move consists of the player flipping over two cards. If they match, the game leaves them face up. Otherwise, they are flipped back face down. The goal is to flip all the cards face up in the minimum number of moves. Figure 1b shows a screenshot of a completed Memory assignment.

## 2.2 Peer Grading

Since submissions in IIPP are all interactive programs, they are very difficult to grade automatically, so IIPP utilizes Coursera's peer grading facilities.

Peer grading takes place after all students submit their programs. Students are required to assess five of their peers' programs and their own program. For each program, grading instructions are presented to the students followed by a series of rubric items. Figure 3 shows the rubric supplied to students for Stopwatch.

For each rubric item, students need to choose how many points to assign using a drop-down menu. After choosing a number of points from the drop-down menu, students can provide additional comments for that specific rubric item in a text area right below the

drop-down menu. After grading the program, students can optionally provide a comment giving overall feedback on the submission.

**Peer Grading Quality.** We estimate that it takes an average grader about 10 minutes to do a reasonable job of grading a typical IIPP assignment, meaning that the grading load for each student is one hour per assignment. In our experience, some students voluntarily spend more time than that grading, but unfortunately, some students spend much less time. In fact, the amount of time and effort spent grading—and hence the accuracy of the peer grading effort as well as its utility to the students being graded—varies widely.

For just one example of this variance, consider Figure 4 which shows an average to above-average grading effort for one particular Stopwatch submission. The grader took off a few points for four of the eight rubric items (giving eight out of 13 points total), and for each of those items, reasonable comments were given.

In contrast, a second grader for this very same submission gave full credit (13 out of 13) to the submission. No comments at all were offered, except for a terse "Very fancy timer :) Great job." under the "Overall" category. Disparities in grading effort such as this are common, and are precisely our motivation for undertaking the study described in the remainder of the paper. We wish to ask the question: How might we motivate students to put in the effort required to produce an evaluation that is of equivalent quality to the evaluation depicted in Figure 4?

# 3. GRADING THE GRADERS

At the highest level, the approach we explore to motivating peer graders is to expose the peer graders themselves to grading, and ask them to grade others' grading efforts. Our hypothesis is that if graders know that they will be evaluated, and if they see the sort of grading efforts put forth by their peers, they will be more motivated to submit high quality evaluations.

In this section, we briefly describe the relatively simple software infrastructure that we implemented to evaluate this idea.

## 3.1 Stage One: Grading the Graders

A few days after students participating in the IIPP "grading the graders" regime finish their peer evaluations, they receive an email with a link to a web application. Following this link leads the student to a page where they are asked to evaluate a set of peer evaluations for five random submissions.

Following this link presents a simple web page that has a link to the assignment that is the subject of the various evaluations. A screen shot of the web page is shown in Figure 5. The web application allows students to cycle through the various rubric items, one at a time. For each rubric item, six different peer evaluations are shown. The application displays the score assigned by each of the six peer evaluators, as well as any comments. The student who is evaluating the evaluators is instructed to click a radio button next to each of the evaluations. The buttons are labeled with "good," "neutral," and "bad," referring to the quality of the evaluation. Typically, a student will look through the various evaluations and comments, and if the evaluators all give the rubric item full credit, the student will assign each a score of "good." If one of the evaluators has taken off some credit, the student will look at the submitted assignment to see if he/she agrees with the loss of credit, and evaluate each of the evaluations accordingly.

After the student evaluates each of the evaluations, the student clicks the "submit" button to move onto the next rubric item. After cycling through each of the rubric items, a final web page informs the student that he/she has completed the "grading the graders" process.

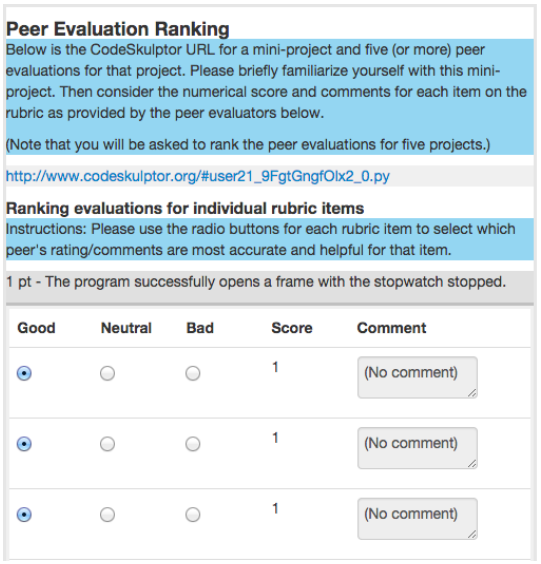



Figure 5: Screen shot of web page that students use to evaluate peer evaluations. A link to a particular submitted assignment is given, along with a listing of five different peer evaluations for that assignment, for a particular rubric item.

## 3.2 Stage Two: Examining Evaluations

Some time later, students who complete the "grading the graders" process will receive an email with a link to a web application that allows the student to see how others evaluated his or her *own* peer grading efforts. Since students are required to evaluate six assignments (five other students' assignments, as well as their own), following the link presents a web page that first lists links to six assignments that the student evaluated. For a particular rubric item, the web page lists each of the six evaluations that the student performed, along with the ratings supplied by those who participated in the "grading the graders" activity. Figure 6 shows a screenshot of the interface.

After cycling through the various rubric items, students are then given the opportunity to examine the grading of their own assignments. That is, the student can see how those who "graded the grader" viewed the quality of the grades for his or her own submission on the assignment. The student can choose to go to a screen that lists (for a particular rubric item) all of the evaluations of his or her own submission, along with the number of times that an examiner determined that the evaluation was "good," "neutral," or "poor."

# 4. EXPERIMENTAL DESIGN

Our central goal is to evaluate whether or not such a framework might have some utility in motivating students to perform high-quality peer evaluations. In this section, we describe in detail the study that we designed and executed to this effect.

## 4.1 Study Overview

The study was open to all participants in the Fall 2013 incarnation of the IIPP class. Because we would be asking participants to do a non-trivial amount of work, and (more importantly) because participants would have their work examined by others, participation needed to be voluntary. There was some concern that voluntary participation would skew the results, making it more difficult to detect effects. After all, those MOOC participants who are motivated

| Item | Points | Description |
|---|---|---|
| 1 | 1 pt | The program successfully opens a frame with the stopwatch stopped. |
| 2 | 1 pt | The program has a working "Start" button that starts the timer. |
| 3 | 1 pt | The program has a working "Stop" button that stops the timer. |
| 4 | 1 pt | The program has a working "Reset" button that stops the timer (if running) and resets the timer to 0. |
| 5 | 4 pt | The time is formatted according to the description in step 4 above. Award partial credit corresponding to 1 pt per correct digit. For example, a version that just draws tenths of seconds as a whole number should receive 1 pt. A version that draws the time with a correctly placed decimal point (but no leading zeros) only should receive 2 pts. A version that draws minutes, seconds and tenths of seconds but fails to always allocate two digits to seconds should receive 3 pts. |
| 6 | 2 pt | The program correctly draws the number of successful stops at a whole second versus the total number of stops. Give one point for each number displayed. If the score is correctly reported as a percentage instead, give only one point. |
| 7 | 2 pt | The "Stop" button correctly updates these success/attempts numbers. Give only one point if hitting the "Stop" button changes these numbers when the timer is already stopped. |
| 8 | 1 pt | The "Reset" button clears the success/attempts numbers. |

Figure 3: Rubric given to peer graders for Stopwatch program.

| Rubric Item | Score | Comments |
|---|---|---|
| 1 | 1 | (No Comments) |
| 2 | 1 | (No Comments) |
| 3 | 1 | (No Comments) |
| 4 | 0 | you forgot use variable timer to stop the timer at reset button. |
| 5 | 1 | You used non-decimal number to count. The numbers for A, C and D was 0-9 and B was 0 - 6.The function format did not pass to the test numbers. |
| 6 | 1 | In the test, i stoped the clock 2.9 and the program showed 2.0 and count 1 correct attempt.I think some problems occurred because the way how you count the number with a non-integer number. |
| 7 | 2 | Update the numbers, but you need to look more deep how your time is increasing because sometimes the clock stop at time but it is not the real time counted. Just put a print time_elapsed at timer_handler an you will can see that behavior |
| 8 | 1 | (No Comments) |
| Overall | N/A | Remember to look more carefully to all the section "Mini-project development process". I think almost of all problems came from count time with a non-integer.Review format function an test it : http://www.codeskulptor.org/#examples-format_template.py. At "Discussion Forum" -> Code Clinic you always find great help to understand some things...use more discussion forum. Sorry about my poor english. =[ google translator help me a lot hehehe |

Figure 4: Reasonable quality peer grading effort for a Stopwatch submission. This grader gave eight out of 13 possible points for the submission. In contrast, for the same submission, another grader gave full credit, and the only comment offered was "Very fancy timer :) Great job." as an "Overall" comment.

enough to participate in a study are probably far more likely to already be motivated enough to submit high quality peer evaluations without the extra incentive (possibly) provided by a "grading the graders" regime. However, since the bias was far more likely to result in dampening of the significance of the study results (rather than creating positive results when none should have existed), in the end we considered this a necessary evil that we could live with.

Study participants were divided into three groups:

1. *Those receiving the full "grading the graders" treatment*. These participants evaluate other peer evaluations (as described in Section 3.1), and have their own peer evaluations evaluated. Then they are asked to examine the evaluations of the peer evaluations that they have performed (as described in Section 3.2), and hence receive the full treatment described in the previous section of the paper. We call this group $G_1$.

2. *Those who only have their peer examinations evaluated*. These participants do not actually evaluate any other peer evaluators, but they have their own evaluations examined by members of $G_1$. We included this group to try to understand whether there is a difference between being asked to evaluate others as is the case in $G_1$ (which necessarily imparts some knowledge of community standards in peer grading to the examiner) and being motivated by knowing that others will be examining one's peer evaluations. We call this group $G_2$.

3. *The control group*. These are people who sign up for the study, but then are not asked to do anything other than participate as usual in the IIPP class. We call this group $G_3$.

By design, we set the ratio of the sizes of three groups to be $G_1 : G_2 : G_3 = 8 : 1 : 1$. The reason for the large size of $G_1$ is that we needed the group to be large enough that it could produce enough evaluations so that members of $G_2$ could consume evaluations of their peer grading efforts, without contributing any evaluations of their own. We were concerned about the effect group size imbalance might have on the statistical power of any analyses that we would need to run, but again, this seemed necessary.

In order to enroll in the study, students were asked to complete a simple web consent form and submit their email address. The consent form described that there were three groups that students could be assigned to (including one where they would be asked to do nothing more than they would normally do in the IIPP class), but not the specific study goals nor what was being measured. Because students could easily communicate on the Coursera forums,

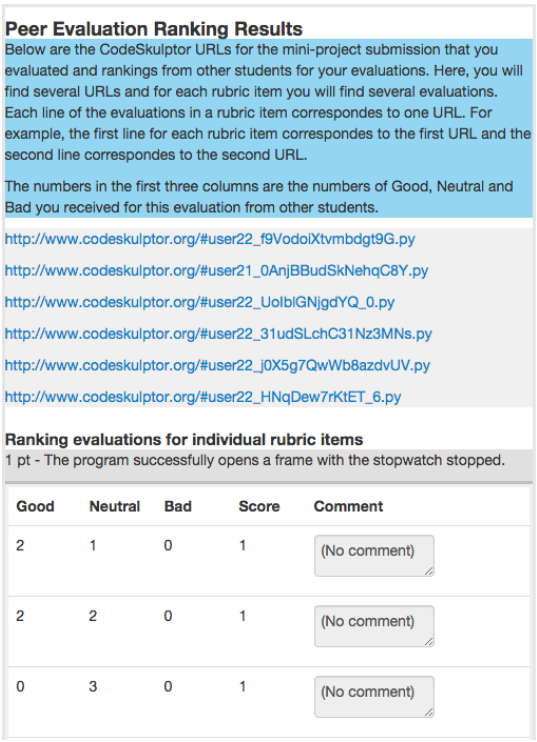



Figure 6: Screen shot of web page that allows students to see how others judged the quality of their peer evaluations.

there was little point in trying to blind the study so that students would not understand what group they were in. In fact, we decided that attempting to blind the study in this way would be worse than not, since we were concerned that it would encourage students to carefully compare notes on class forums regarding what they were seeing, possibly biasing the results.

To motivate students to sign up for the study—we were especially interested in attracting somewhat less motivated students from the general student population who might not have otherwise signed up—a Nexus tablet was promised to one randomly-selected student from each group.

## 4.2 Timeline and Details

All students who submitted the consent form were randomly assigned to groups. 3,015 students completed a consent form during the enrolling phase. 2,412 students were assigned to $G_1$, 301 students to $G_2$ and 302 students to $G_3$. All students received an email with information describing what they needed to do in the study.

It was expected that all students in the study would participate in the study during both the Stopwatch and Memory assignments. Here we detail the timeline that was used for both assignments.

**Day 1.** The assignment (Stopwatch/Memory) is posted and a special submission page for the study (separate from the normal submission page) is opened; study participants were asked to submit to that particular page. Students had nine days to submit.

**Day 9.** The assignment ends. If a study participant did not submit to the special submission page by day 9, he or she was removed from the study. Students begin peer evaluation. They have one week.

**Day 16.** Peer evaluation ends. At this point, the evaluation phase begins and emails are sent to all $G_1$ students pointing them to the web page where they can evaluate others' evaluations. Three days were allotted to this task. 201 students outside of the study (in the case of Stopwatch) and 421 students outside of the study (in the case of Memory) also (mistakenly) submitted to the special study submission page. We were happy to have extra evaluations to work with, so these students were treated as $G_1$ students and asked to "grade the graders," but they are not otherwise included in the study results. (Interestingly, 60 of the 201 "mistaken Stopwatch" students and 124 of the 421 "mistaken Memory" students actually completed the "grading the graders" task of Section 3.1).

**Day 19.** "Grading the graders" ends. In the case of Stopwatch, 1,891 out of 2,412 $G_1$ students and 244 out of 301 $G_2$ students (those who had successfully completed study requirements) receive emails pointing them to a URL where they can see what others thought of their submitted evaluations. In the case of Memory, 1,387 $G_1$ students and 192 $G_2$ students receive this email.

# 5. EXPERIMENTAL RESULTS

In this section, we describe in detail the results we obtained by analyzing the data that we collected.

## 5.1 Hypotheses Tested

Our goal was to determine whether or not there was some evidence that students enrolled in $G_1$ did a better job grading compared to either $G_2$ or $G_3$, or both. As we will discuss in detail shortly, if we find that $G_1$ generally does a better job, then it might be taken as evidence that the "grading the graders" regime does in fact motivate peer graders.

Thus, our data analysis task comes down to measuring the quality of students' grading efforts. We felt that (short of having a human expert review on the order of 10,000 peer grading submissions) the two best proxies for measuring quality are (1) whether or not an evaluator gets it right, and typically gives a high score to a good program and typically gives a low score to a bad program, and (2) how much time and effort are taken in writing comments.

However, it turns out that it is not trivial to measure either of those proxies. The problem with looking at score accuracy is that we do not actually know when a submission is in fact a good program, and when it is bad. These are, after all, interactive programs that are very difficult to grade automatically. This is why the IIPP course relies on peer grading. And without actually reading the comments, it is difficult to measure the effort level.

To address the first difficulty, we manually devised a number of assignment-specific program analyses that were able to automatically and roughly categorize student submissions as being good or bad. These methods simulated each submission on a large number of manually written "tests" — defined here as finite sequences of events — and recorded the program's executions on these tests. A large number of manually designed rules were then used to judge the correctness of these executions. The analyses that we ran were not perfect. However, we used deep knowledge about the assignments, as well as a large amount of effort, to make our tests and rules as accurate as possible. Moreover, any inaccuracy in the classification will tend to mask the accuracy (or inaccuracy) of a set of program grades in a systematic way (since a mis-categorized program will be graded by members of $G_1$, $G_2$ and $G_3$) and so such inaccuracies are unlikely to introduce bias into our analysis.

To address the second difficulty, we decided to use comment length as a reasonable measure of the quality of a comment. Of course, it is always possible for a grader to write a long but inane and useless comment, but in general, one would expect a longer comment to correlate with more care on the part of the grader.

| | | Stopwatch | | | Memory | | |
|---|---|---|---|---|---|---|---|

$H_0^1$ = "The mean score for group $A$ is no greater than the mean score for group $B$ on good programs"

| Group $A$ | Group $B$ | Group $A$ Mean | Group $B$ Mean | $p$-value | Group $A$ Mean | Group $B$ Mean | $p$-value |
|---|---|---|---|---|---|---|---|
| $G_1$ | $G_2$ | 12.91 | 12.91 | 0.5889 | 10.88 | 10.89 | 0.7781 |
| $G_1$ | $G_3$ | 12.91 | 12.88 | **0.0255** | 10.88 | 10.87 | 0.3501 |
| $G_2$ | $G_3$ | 12.91 | 12.88 | **0.0477** | 10.89 | 10.87 | 0.1987 |

$H_0^2$ = "The mean score for group $A$ is no less than the mean score for group $B$ on bad programs"

| Group $A$ | Group $B$ | Group $A$ Mean | Group $B$ Mean | $p$-value | Group $A$ Mean | Group $B$ Mean | $p$-value |
|---|---|---|---|---|---|---|---|
| $G_1$ | $G_2$ | 12.02 | 12.09 | 0.1095 | 10.29 | 10.36 | 0.1911 |
| $G_1$ | $G_3$ | 12.02 | 12.04 | 0.3543 | 10.29 | 10.39 | 0.0783 |
| $G_2$ | $G_3$ | 12.09 | 12.04 | 0.7527 | 10.36 | 10.39 | 0.3214 |

$H_0^3$ = "The median comment length for group $A$ is no greater than the median comment length for group $B$ on good programs"

| Group $A$ | Group $B$ | Group $A$ Median | Group $B$ Median | $p$-value | Group $A$ Median | Group $B$ Median | $p$-value |
|---|---|---|---|---|---|---|---|
| $G_1$ | $G_2$ | 11 | 10 | 0.6603 | 12 | 10 | **0.0060** |
| $G_1$ | $G_3$ | 11 | 11 | 0.7888 | 12 | 10 | **0.0036** |
| $G_2$ | $G_3$ | 10 | 11 | 0.8341 | 10 | 10 | 0.9356 |

$H_0^4$ = "The median comment length for group $A$ is no greater than the median comment length for group $B$ on bad programs"

| Group $A$ | Group $B$ | Group $A$ Median | Group $B$ Median | $p$-value | Group $A$ Median | Group $B$ Median | $p$-value |
|---|---|---|---|---|---|---|---|
| $G_1$ | $G_2$ | 83 | 78 | 0.3300 | 69 | 51.5 | 0.1014 |
| $G_1$ | $G_3$ | 83 | 74 | 0.2166 | 69 | 51 | 0.1348 |
| $G_2$ | $G_3$ | 78 | 74 | 0.4077 | 51.5 | 51 | 0.4700 |

$H_0^5$ = "The fraction of people doing a 'bad job' in group $A$ is no less than the fraction doing a 'bad job' in group $B$"

| Group $A$ | Group $B$ | Group $A$ Mean | Group $B$ Mean | $p$-value | Group $A$ Mean | Group $B$ Mean | $p$-value |
|---|---|---|---|---|---|---|---|
| $G_1$ | $G_2$ | 0.3820 | 0.4007 | 0.0721 | 0.3477 | 0.4021 | **0.0004** |
| $G_1$ | $G_3$ | 0.3820 | 0.4091 | **0.0191** | 0.3477 | 0.4157 | **0.0001** |
| $G_2$ | $G_3$ | 0.4007 | 0.4091 | 0.3152 | 0.4021 | 0.4157 | 0.2504 |

$H_0^6$ = "The fraction of people doing a 'really bad job' in group $A$ is no less than the fraction doing a 'really bad job' in group $B$"

| Group $A$ | Group $B$ | Group $A$ Mean | Group $B$ Mean | $p$-value | Group $A$ Mean | Group $B$ Mean | $p$-value |
|---|---|---|---|---|---|---|---|
| $G_1$ | $G_2$ | 0.04522 | 0.05051 | 0.1894 | 0.03404 | 0.04442 | 0.0691 |
| $G_1$ | $G_3$ | 0.04522 | 0.06133 | **0.0057** | 0.03404 | 0.04648 | **0.0400** |
| $G_2$ | $G_3$ | 0.05051 | 0.06133 | 0.0792 | 0.04442 | 0.04648 | 0.3896 |

Figure 7: Summary of study results, comparing $G_1$ (students who both "graded the graders" and viewed the peer evaluations of their own grading efforts), $G_2$ (students who only viewed the peer evaluations of their own grading efforts), and $G_3$ (students who neither "graded the graders" nor viewed the peer evaluations of their own grading efforts). Findings significant at the $0.05$ level are shown in **bold**.

With this in mind, we developed six different null hypotheses that we would test in an attempt to differentiate the quality of any two groups of graders $A$ and $B$. These null hypotheses are:

**Hypothesis One:** $H_0^1$ = "The mean score for group $A$ is no greater than the mean score for group $B$ on good programs"

If this hypothesis is refuted, then it means that group $A$ is doing a better job than group $B$ in recognizing good programs, which would be a strong indicator that group $A$ does a better job grading good programs.

**Hypothesis Two:** $H_0^2$ = "The mean score for group $A$ is no less than the mean score for group $B$ on bad programs"

If this is refuted, it means that group $A$ does a better job than $B$ recognizing bad programs, which is again indicative that group $A$ is doing a better job.

**Hypothesis Three:** $H_0^3$ = "The median comment length for group $A$ is no greater than the median comment length for group $B$ on good programs"

If this is refuted, it means that group $A$ writes longer comments on high-quality submissions than group $B$. We use median rather than mean since the comment length appears to have a heavy-tailed distribution, making the mean quite unstable. Refuting this would be particularly interesting, because one might expect that it is very easy for a mediocre grader to simply give full credit to a high-quality submission. A careful grader would look at the code, even if the program works well, and offer comments on the style and substance of the implementation.

**Hypothesis Four:** $H_0^4$ = "The median comment length for group $A$ is no greater than the median comment length for group $B$ on bad programs"

Comments on bad programs are the most important feedback that a struggling student will receive, and so this is also an important hypothesis.

**Hypothesis Five:** $H_0^5$ = "The fraction of people doing a 'bad job' in group $A$ is no less than the fraction doing a 'bad job' in $B$"

We define someone who has done a "bad job" to be a grader that either (a) gets the grade wrong, and gives a perfect score to

a program that our code analysis engine thinks is flawed (or gives a non-perfect score to a program that our engine can find no fault with), or (b) writes no comment across all rubric items. If we are able to refute this hypothesis, it means that someone from group $A$ is less likely to do a bad job than someone from group $B$.

**Hypothesis Six:** $H_0^6$ = "The fraction of people doing a 'really bad job' in group $A$ is no less than the fraction doing a 'really bad job' in group $B$"

We define someone who has done a "really bad job" similarly to the way we define someone who has done a "bad job," but we replace the *or* with an *and*.

For each of these hypotheses, we perform six different statistical tests. For Stopwatch, we compare (1) $G_1$ vs. $G_2$, (2) $G_1$ vs. $G_3$, and (3) $G_2$ vs. $G_3$. We also make the same comparisons for Memory. The reason that we performed these particular tests is that we were looking for evidence that the "grading the graders" regime has a positive effect on peer grading, and so it makes sense to compare those who have undertaken the full "grading the graders" program (those in $G_1$) versus the other two groups. We are also interested in comparing $G_2$ and $G_3$ because we would like to see if there is any difference between the partial "grading the graders" program (those in $G_2$ only had their peer evaluations evaluated; they did not actually evaluate others' peer evaluations) and the control group.

## 5.2 Statistical Significance

Checking whether these various null hypotheses are refuted obviously requires some sort of statistical test of significance to obtain a $p$-value for each of the hypotheses. At first glance, the textbook test for this sort of experimental setup would be a paired $t$-test [4], since we have two groups, $A$ and $B$, and in each case we are checking for a difference in the mean or the median of some statistic, computed across the groups. Further, a "paired" version of the test seems appropriate, since members of both groups are typically paired, grading the same program.

Unfortunately, a $t$-test, paired or otherwise, seems inappropriate when examined in detail. In fact, any sort of textbook test for significance, parametric or not, is likely going to be invalid in our experimental framework. The problem is that when we test a particular hypothesis (say, $H_0^1$) we have multiple sources of correlation across the scores that are being added. Not only are the graders grading the same programs (a source of correlation that is in fact handled by a paired test) but *more than one grader from each group* may be grading the same program. For example, when looking at a specific program, it can be the case that five graders from group $A$ and three graders from group $B$ graded the same program, leading to a very unique covariance structure. Another source of correlation among the observed scores is that *each grader will grade multiple programs*, so that the scores may be correlated because they came from the same grader.

As a result, we had two obvious options. We could resort to something like a $t$-test, being cognizant of its limitations, or else we could utilize a simulation-based solution that naturally takes into account such issues, such as the bootstrap [13]. In the end, we chose the latter option.

Briefly, the idea behind the bootstrap is to use a resampling-based algorithm to simulate a very large number of data sets from the collected data set. The null hypothesis is checked on each, and the fraction of the time that the null hypothesis holds is the $p$-value.

To apply the bootstrap in our own setting, we generate a simulated data set as follows. Given a set of $n$ programs called $P$ for which one or more peer graders participating in the study actually graded, let $cnt(A, p)$ denote the number of graders from group $A$ who graded program $p \in P$, and let $scores(A, p)$ denote the set of peer grading efforts created by those graders from group $A$. Then to bootstrap resample our data set, we first resample $n$ programs from $P$ by sampling $n$ times from $P$ with replacement. Call these sampled programs $p_1, p_2, ..., p_n$. For each $p_i$, we then create a new set of grading efforts for group $A$ by resampling $cnt(A, p_i)$ grades from $scores(A, p_i)$. We then create a new set of grading efforts for group $B$ by resampling $cnt(B, p_i)$ grades from $scores(B, p_i)$. By unioning all of the grading efforts across all $p_1, ..., p_n$, we create a new, simulated version of our data set. This simulated version respects the correlations induced by having multiple graders grade the same program (because there will be multiple grades of the same program in the simulated data set), and it also respects correlations induced by having the same grader grade multiple programs (since this will also happen in the simulated data set).

## 5.3 Results

We ran the resulting bootstrap tests across all of the hypotheses defined above. All results are given above in Figure 7. For each group and for each hypothesis, we give the $p$-value with which we reject the relevant null hypothesis, according to the bootstrap test. In general, a $p$-value of less than or equal to $0.05$ (or possibly $0.1$) is considered to be statistically significant. We **bold** all $p$-values that are significant at $\leq 0.05$. Just as important, we give the mean or median value of the relevant statistic for each group that is being tested. For example, for $H_0^1$ = "The mean score for group $A$ is no greater than the mean score for group $B$ on good programs," we give the mean program score for good programs for group $A$ and group $B$. These values are there to let the reader judge whether any differences are of practical significance.

In the remainder of the section, we highlight and explain a few of the results. In the next full section of the paper, we discuss the conclusions that we might draw from them.

**Many of the Findings Are Statistically Significant.** If, for a moment, we restrict ourselves to comparisons of $G_1$ vs. $G_2$ and $G_1$ vs. $G_3$, for Memory (which is the second assignment; hence, any gains from undertaking the "grading the graders" regime on Stopwatch would have had a chance to manifest themselves), 12 different null hypotheses were checked. Of those 12, 5 were rejected with a $p$-value $\leq 0.05$, and the other 2 at a $p$-value $\leq 0.1$. While there is certainly something of a multiple-hypothesis testing problem here given that 12 tests were run [23], the fact so many result in rejection of the null hypothesis seems strongly indicative of a positive effect of the full "grading the graders" regime compared to students in $G_2$ and the control group $G_3$.

**Might We Have Hurt Statistical Power By Partitioning $\bar{G_1}$ into $G_2$ and $G_3$?** Often, if the effect one is looking for is stronger in one segment of the population, it makes sense to stratify into subpopulations and run multiple tests, but since it results in multiple tests that each have a smaller number of samples, power is reduced.

To investigate this a bit, we re-ran each of the hypothesis tests, this time comparing $G_1$ vs. $\bar{G_1}$. The results are summarized in Figure 8. Again restricting ourselves to Memory, 3 of 6 null hypotheses are rejected with a $p$-value $\leq 0.02$, and another 2 of 6 with a $p$-value $\leq 0.056$.

**No Significant Difference Between $G_2$ and $G_3$.** Of the twelve comparisons between $G_2$ and $G_3$ shown in Figure 7, only one rejects the null hypothesis at a $p$-value $\leq 0.05$. This seems to be strong evidence for the position that simply knowing that one's grading efforts are going to be examined does not help to increase grading quality—an issue we will consider subsequently.

| | | Stopwatch | | | Memory | | |
|---|---|---|---|---|---|---|---|

$H_0^1$ = "The mean score for group $A$ is no greater than the mean score for group $B$ on good programs"

| Group $A$ | Group $B$ | Group $A$ Mean | Group $B$ Mean | $p$-value | Group $A$ Mean | Group $B$ Mean | $p$-value |
|---|---|---|---|---|---|---|---|
| $G_1$ | $G_2$ and $G_3$ | 12.912094 | 12.898951 | 0.1305 | 10.875696 | 10.878788 | 0.5824 |

$H_0^2$ = "The mean score for group $A$ is no less than the mean score for group $B$ on bad programs"

| Group $A$ | Group $B$ | Group $A$ Mean | Group $B$ Mean | $p$-value | Group $A$ Mean | Group $B$ Mean | $p$-value |
|---|---|---|---|---|---|---|---|
| $G_1$ | $G_2$ and $G_3$ | 12.021513 | 12.069231 | 0.1610 | 10.286411 | 10.373171 | 0.0557 |

$H_0^3$ = "The median comment length for group $A$ is no greater than the median comment length for group $B$ on good programs"

| Group $A$ | Group $B$ | Group $A$ Median | Group $B$ Median | $p$-value | Group $A$ Median | Group $B$ Median | $p$-value |
|---|---|---|---|---|---|---|---|
| $G_1$ | $G_2$ and $G_3$ | 11 | 10 | 0.7088 | 12 | 10 | **0.0009** |

$H_0^4$ = "The median comment length for group $A$ is no greater than the median comment length for group $B$ on bad programs"

| Group $A$ | Group $B$ | Group $A$ Median | Group $B$ Median | $p$-value | Group $A$ Median | Group $B$ Median | $p$-value |
|---|---|---|---|---|---|---|---|
| $G_1$ | $G_2$ and $G_3$ | 83 | 76.5 | 0.2111 | 69 | 51.5 | 0.0540 |

$H_0^5$ = "The fraction of people doing a 'bad job' in group $A$ is no less than the fraction doing a 'bad job' in group $B$"

| Group $A$ | Group $B$ | Group $A$ Mean | Group $B$ Mean | $p$-value | Group $A$ Mean | Group $B$ Mean | $p$-value |
|---|---|---|---|---|---|---|---|
| $G_1$ | $G_2$ and $G_3$ | 0.382037 | 0.40477 | **0.0132** | 0.347659 | 0.408338 | **0** |

$H_0^6$ = "The fraction of people doing a 'really bad job' in group $A$ is no less than the fraction doing a 'really bad job' in group $B$"

| Group $A$ | Group $B$ | Group $A$ Mean | Group $B$ Mean | $p$-value | Group $A$ Mean | Group $B$ Mean | $p$-value |
|---|---|---|---|---|---|---|---|
| $G_1$ | $G_2$ and $G_3$ | 0.045217 | 0.05577 | **0.0087** | 0.034042 | 0.045371 | **0.0182** |

Figure 8: Summary of study results, comparing $G_1$ (students who both "graded the graders" and viewed the peer evaluations of their own grading efforts) versus those who did not "grade the graders". Findings significant at the $0.05$ level are shown in **bold**.

**Many of the Findings Are of Practical Significance.** It is often easy to conflate statistical significance with practical significance. When one runs a large-scale study involving thousands of participants, there are often statistically significant results that are of no practical significance. Is that the case here? Many of the results in Figure 7 and Figure 8 appear to be of practical significance as well. For one example, consider the Memory assignment results in Figure 8. We found that the percentage of graders who did a "really bad job" decreased from 4.5% for those who do not receive the full "grading the graders" treatment down to 3.4% for those who did. The percentage of those who do a "bad job" decreased from 41% to 35% under the regime. These reductions are perhaps more significant when one considers that the students who voluntarily signed up for the study are likely to be far more motivated already than those who did not.[3]

**The Effects Were Lasting.** Students completed three more assignments after the study ended. To see whether there were any lasting effects from the "grading the graders" regime, we analyzed the median comment lengths of student grading efforts on those assignments (as a sanity check, we also analyzed the comment lengths on the assignment immediately before Stopwatch). Figure 9 summarizes the results. We find that there is actually a significant, persistent effect of participating in the "grading the graders" regime.

# 6. CONCLUSIONS

It is easy to do a poor job peer grading. Intuitively, if a grader spends little time grading, then the grader will not find any problems. This leads to artificially inflated numerical grades and little feedback. As intimated in the introduction of the paper, we found that it was much more likely that a grader would give a perfect score to an imperfect program than the other way around, which supports this intuition. The students who did not do well on an assignment and most need quality feedback are the ones who suffer the most from poor grading.

Our study results corroborate the expectation that peer graders have no problem assigning high numerical scores to good assignments. There was little evidence of a difference among study groups at assigning numerical grades to good IIPP programs. While there were statistically significant results showing that both groups $G_1$ and $G_2$ did a better job at assigning a grade to good Stopwatch programs, the actual differences in scores among all the groups for both Stopwatch and Memory were in the hundredths of points. Further note that the Stopwatch program was actually simpler to grade, as Memory requires a more complex set of actions to be performed by the grader to verify correctness. More than anything, this indicates that all groups graded good programs well.

Ultimately, this suggests that with no intervention, it should be expected that peer grading will bias towards higher scores. However, while little effort is required to give high scores, effort is required provide feedback. Here we found the "grading the graders" regime to be useful for motivating the graders, even on good programs. We found that group $G_1$ did provide longer comments than both $G_2$ and $G_3$ on such programs.

In fact, the study supports the hypothesis that the full regime leads to better peer grading in general. Those in group $G_1$ *did* consistently do a better job grading, according to most of our metrics. Consider the Stopwatch results in Figure 8, where in $5$ of $6$ analyses, the null hypothesis was rejected with a $p$-value $\leq 0.056$. The only case where the null hypothesis was *not* rejected was when checking whether those in $G_1$ did no better in scoring good programs. But this is not surprising, as discussed above.

[3]Along those lines, we did a bit of data analysis on the students who did not agree to participate in the study but managed to accidentally submit their grading efforts to the study. We found, for example, that the median comment length for everyone who signed up for the study was 47, while for the non-study group the median length was 17. Thus, there is a lot more room to improve the grading efforts of non-study participants.

| | | Before Study | | | After Study | | |
|---|---|---|---|---|---|---|---|
| $H_0$ = "The median comment length for group $A$ is no greater than the median comment length for group $B$ on all programs" | | | | | | | |
| Group $A$ | Group $B$ | Group $A$ Median | Group $B$ Median | $p$-value | Group $A$ Median | Group $B$ Median | $p$-value |
| $G_1$ | $G_2$ | 46 | 56.5 | 0.9472 | 21 | 16 | **0.0437** |
| $G_1$ | $G_3$ | 46 | 47.5 | 0.5650 | 21 | 16 | **0.0184** |
| $G_2$ | $G_3$ | 56.5 | 47.5 | 0.0991 | 16 | 16 | 0.5595 |
| $G_1$ | $G_2$ and $G_3$ | 46 | 52 | 0.8950 | 21 | 16 | **0.0072** |

Figure 9: Analyzing comment lengths in the assignment before the study began (at left) and for the three assignments after the study ended. Findings significant at the 0.05 level are shown in **bold**.

Not only were the results immediately noticeable, they also seemed to be lasting. In Figure 9 we see that the median comment length for those in $G_1$ stays greater for the final three assignments in the class, compared to the other study participants. Significantly, this was after the end of the study, when students had no reason to expect that the median comment length was going to be monitored.

One might easily believe that peer graders can be motivated to do a better job simply by telling them that their evaluations will themselves be evaluated. However, we found no evidence that this is effective. In particular, those in $G_2$ tended to do no better than those in $G_3$ throughout the study. This strongly suggests that knowing that one is being monitored by one's peers is not a strong enough motivation to do a good job grading others' assignments. We suspect that this extends to other forms of monitoring, such as the use of sentinels to catch bad grading.

Another surprising result is that there were some significant differences between $G_1$ and both of the other two groups not only on the Memory assignment, but also on the Stopwatch assignment (see Figure 8). This was surprising because Stopwatch grading took place *after* students agreed to participate in the study (and after they had been assigned to groups), but *before students had actually participated in the study in any meaningful way*. At the point that the data were collected, the participants had yet to actually participate in the protocol described in Sections 3.2 and Section 3.1, so they had not yet examined any other students' grading efforts, and yet the differences between groups were still significant.

The most likely explanation is that those students in $G_1$ were aware that they were participating in the full version of the study and hence they were enthusiastic and felt motivated to do well by this simple fact—they somehow felt "special," which turned out to be highly motivational by itself. The comments on the Coursera forums seem to corroborate this explanation. This observation is one source of our belief that high quality grading is likely as much related to good motivation as it is to good information.

Given all of this, we believe that motivation is a key component of high-quality grading, and that actually seeing cases where other students put in the effort to do a good job (or, conversely, seeing how unhelpful it is when other students do not put in such effort) helps provide motivation to student graders. Thus, we would recommend that MOOCS which utilize peer grading should consider using something like the "grading the graders" regime early on in a class, in order to help train and motivate students to do a good job grading subsequent assignments.

# 7. RELATED WORK

Peer evaluation has a long history in education. Many researchers have demonstrated that peer evaluation can be a beneficial part of the learning process [27, 10, 16]. Further, controlled studies have shown that peer evaluation can be a good proxy for instructor feedback [15, 30]. In addition to lessening the grading burden on instructors, peer evaluation is a valuable learning activity for the peer grader [14]. Grading someone else's work requires retrieval and evaluation of core concepts—activities known to boost learning and retention [11].

In the context of a MOOC, peer evaluation offers some unique challenges compared to a traditional classroom environment. The first is the lack of a gating function. In a traditional classroom, most students will have taken the same preparatory classes and/or been admitted to the same school. This is not the case in a typical MOOC, where one can assume no core competence on the part of the evaluators. A second challenge is the lack of centralized oversight. There is much mention in the online education research community of the necessity of "teaching presence" [25, 28]: the need for participants to feel as if an instructor is present and monitoring the proceedings. In a traditional environment where peer evaluation may be applied, a teacher is present to settle disputes and monitor the fairness of the proceedings. The incentive system that we investigate can be seen as a decentralized, bottom-up simulation of teacher presence.

Student behavior in an online education setting has been studied before. Specifically, Davies and Graff [7] studied correlation between students' performance and the level of interaction between other students. They discovered that greater online interaction did not lead to significantly higher performance, but they did find that students who failed the course tend to interact less.

User performance and behavior patterns have been studied in the context of MOOCs. Breslow et al. [5] did a broad study with the data gathered from edX 6.002x. Their study looked at resource usage, student demographics, achievement, persistence and success. Hew and Cheung [18] discovered several challenges in MOOCs. One of the challenges is low student engagement rate. Anderson et al. [3] examined student behavior patterns and engagement styles in MOOCs. They found some correlations between different engagement styles and student performance. Kizilcec et al. [20] clustered engagement patterns and discovered four prototypical categories of engagement consistently across three different MOOCs.

A recent paper by Kulkarni et al. [21] considers ways to improve grading accuracy in MOOCs. Specifically, The authors consider data on self and peer assessment from two iterations of a MOOC. Their solutions to improving grading accuracy include giving feedback to students about the bias in their peer grading, and using more precise rubric items.

Some online knowledge-sharing forums such as StackOverflow and Y! Answers use incentives to motivate people to contribute. Recognizing contributions by awarding badges has proved effective in motivating contribution [12]. Anderson et al. [2, 3] studied the effect of using badges to motivate participation in a MOOC. They have found that using badges indeed increases the level of engagement.

# 8. REFERENCES


[1] Susan Adams. Is coursera the beginning of the end for traditional higher education? *Higher Education*, 2012.

[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Steering User Behavior with Badges. In *Proceedings of WWW*, pages 95–106, 2013.

[3] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Engaging with Massive Online Courses. In *Proceedings of WWW*, pages 687–698, 2014.

[4] George EP Box, William Gordon Hunter, J Stuart Hunter, et al. *Statistics for experimenters*. John Wiley and sons New York, 1978.

[5] Lori B. Breslow, David E. Pritchard, Jennifer DeBoer, Glenda S. Stump, Andrew D. Ho, and Daniel T. Seaton. Studying learning in the worldwide classroom: Research into edX's first mooc. *Research & Practice in Assessment*, 8:13–25, 2013.

[6] D. Casey, E. Burke, C. Houghton, L. Mee, R. Smith, D. Van Der Putten, H. Bradley, and M. Folan. Use of peer assessment as a student engagement strategy in nurse education. *Nursing & Health Sciences*, 13(4):514–520, 2011.

[7] Jo Davies and Martin Graff. Performance in e-learning: online participation and student grades. *British Journal of Educational Technology*, 36(4):657–663, July 2005.

[8] J.P. Dineen, H.B. Clark, and T.R. Risley. Peer tutoring among elementary students: Educational benefits to the tutor. *Journal of Applied Behavior Analysis*, 10(2):231, 1977.

[9] A. DiPardo and S.W. Freedman. Peer response groups in the writing classroom: Theoretic foundations and new directions. *Review of Educational Research*, 58(2):119–149, 1988.

[10] F. Dochy, M. Segers, and D. Sluijsmans. The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education*, 24(3):331–350, 1999.

[11] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 2012.

[12] David Easley and Arpita Ghosh. Incentives, Gamification, and Game Theory: An Economic Approach to Badge Design. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, EC '13, pages 359–376, New York, NY, USA, 2013. ACM.

[13] Bradley Efron and B Efron. *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM, 1982.

[14] P.A. Ertmer, J.C. Richardson, B. Belland, D. Camin, P. Connolly, G. Coulthard, K. Lei, and C. Mong. Using peer feedback to enhance the quality of student online postings: An exploratory study. *Journal of Computer-Mediated Communication*, 12(2):412–433, 2007.

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

[16] Nancy Falchikov. Peer feedback marking: Developing peer assessment. *Innovations in Education and Training International*, 32(2):175–187, 1995.

[17] Lisa E. Gueldenzoph and Gary L. May. Collaborative peer evaluation: Best practices for group member assessments. *Business Communication Quarterly*, 65(1):9–20, 2002.

[18] Khe Foon Hew and Wing Sum Cheung. StudentsâĂŹ and instructorsâĂŹ use of massive open online courses (moocs): Motivations and challenges. *Educational Research Review*, 12(0):45 – 58, 2014.

[19] James A Keaten and M. Elizabeth Richardson. A field investigation of peer assessment as part of the student group grading process, 1993.

[20] René F. Kizilcec, Chris Piech, and Emily Schneider. Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. In *Proceedings of the International Conference on Learning Analytics and Knowledge*, LAK '13, pages 170–179. ACM, 2013.

[21] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact.*, 20(6):33:1–33:31, December 2013.

[22] Y. Lai. Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology*, 41(3):432–454, 2009.

[23] Rupert G Miller. *Simultaneous statistical inference*, volume 196. Springer, 1966.

[24] Laura Pappano. The year of the mooc. *The New York Times*, 2(12):2012, 2012.

[25] J.C. Richardson and K. Swan. *Examing social presence in online courses in relation to students' perceived learning and satisfaction*. 2003.

[26] D. Sluijsmans, F. Dochy, and G. Moerkerke. Creating a learning environment by using self-, peer- and co-assessment. *Learning Environments Research*, 1(3):293–319, 1998.

[27] Hugh Somervell. Issues in assessment, enterprise and higher education: the case for self-peer and collaborative assessment. *Assessment and Evaluation in Higher Education*, 18(3):221–233, 1993.

[28] K. Swan and L.F. Shih. On the nature and development of social presence in online course discussions. *Journal of Asynchronous Learning Networks*, 9(3):115–136, 2005.

[29] Terry Tang, Scott Rixner, and Joe Warren. An environment for learning interactive programming. In *Proceedings of the 45th ACM technical symposium on Computer science education*, pages 671–676. ACM, 2014.

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