# Dynamics of Peer Grading: An Empirical Study

TECHNICAL REPORT UCSC-SOE-16-04
SCHOOL OF ENGINEERING
UNIVERSITY OF CALIFORNIA, SANTA CRUZ

Luca de Alfaro*
University of California, Santa Cruz
Department of Computer Science
luca@ucsc.edu

Michael Shavlovsky
University of California, Santa Cruz
Department of Computer Science
mshavlov@soe.ucsc.edu

## ABSTRACT

Peer grading is widely used in MOOCs and in standard university settings. The quality of grades obtained via peer grading is essential for the educational process. In this work, we study the factors that influence errors in peer grading. We analyze 288 assignments with 25,633 submissions and 113,169 reviews conducted with CrowdGrader, a web based peer grading tool. First, we found that large grading errors are generally more closely correlated with hard-to-grade submission, rather than with imprecise students. Second, we detected a weak correlation between review accuracy and student proficiency, as measured by the quality of the student's own work. Third, we found little correlation between review accuracy and the time it took to perform the review, or how late in the review period the review was performed. Finally, we found a clear evidence of tit-for-tat behavior when students give feedback on the reviews they received. We conclude with remarks on how these data can lead to improvements in peer-grading tools.

## 1. INTRODUCTION

In peer grading, students review and grade each other's work. The grades assigned by the students to each item are then merged into a single *consensus grade* for the item. Peer grading has several benefits, as reported in the literature, including the fact that students learn from each other's work, and the reduced workload on the instructors. For these reasons, peer grading has been widely used both in MOOCs, where it would be infeasible for a small number of instructors to grade all work [13, 1, 5, 11], and in standard university classes [16, 14, 9, 17, 3, 15].

Successful peer grading is predicated on the ability to reconstruct a reasonably accurate consensus grade from the grades assigned by the students. This leads to the following question: what factors cause or influence the errors in peer-assigned grades? We are interested in this question for three reasons. First, we wish to obtain a better understanding of the dynamics and human factors in peer grading. Second, a better understanding of the causes of error has the potential to lead to tool improvements that reduce the errors. For example, if mis-understanding on the work submitted constituted a large source of error, then peer grading tools could be augmented with means for work authors and graders to communicate, so that the misunderstandings could be resolved. Third, a better model of peer grading errors might lead to better algorithms for aggregating the student-assigned grades into the consensus grades for each item.

Our interest in the origin of peer-grading errors is also due to our work on the peer-grading tool CrowdGrader[1] [8]. We have put considerable effort in reducing the error in the consensus grade computed by CrowdGrader, as compared to control instructor-assigned grades. While efforts on the tool UI and UX paid off, as we will detail later, the efforts to create more precise grade-aggregation algorithms did not. In the context of MOOCs, [13] reports a 30% decrease in error using parameter-estimation algorithms that infer, and correct for, the imprecision and biases of individual users. CrowdGrader is used mostly in universities and high-schools. On CrowdGrader data, the parameter-estimation algorithm of [13] offers no benefit compared with the simple "Olympic average" obtained by removing lowest and highest grades, and averaging the rest. Indeed, we have spent a large amount of time experimenting with variations upon the algorithm (see also [7]) and new ideas, but we are yet to find an algorithm that offers consistent error reduction of more than 10% compared to the Olympic average. Thus our interest on the origin of errors in CrowdGrader: what are the main causes? What makes them so difficult to remove using algorithms based on parameter estimation, reputation systems, and more?

To gain an understanding of the dynamics of peer grading, we have analyzed a set of CrowdGrader data consisting in 288 assignments, 25,633 submissions, and 113,169 grades and reviews. Of the 25,633 submissions, 2,564 were graded by the instructors in addition to the students. The questions we ask include the following.

*Is error mostly due to items or to students?* We first ask the question of whether the imprecision in peer grades can be best explained in terms of students being imprecise, or items being difficult to grade. We answer this question in two different ways.

*In alphabetical order

[1] www.crowdgrader.org

First, we build a parameterized probabilistic model of the review process, similar to the model of [13], in which every review error is the sum of a component due to the submission being reviewed, and of a component due to the reviewer. The parameters of the model are then estimated via Gibbs sampling [10]. The results indicate that students contribute roughly two thirds of the total evaluation error.

This result, however, speaks to the *average* source of error. Of particular concern in peer grading are the very large errors that happen less frequently, but have more impact on the perceived fairness and effectiveness of peer grading. We measure the correlation of large errors in items, and in users; our results indicate that hard-to-grade items are a more common cause of large errors than very imprecise students.

*Do better students make better graders?* A natural question is whether better students make better graders. In Section 6 we give an affirmative answer: students whose submissions are in the lower 30%-percentile quality-wise have a grading error that is about 15% above average. The effect is fairly weak, a likely testament to the fundamental homogeneity in abilities in a high-school or college class, as well as to the fact that grading a homework is usually easier than solving the homework.

*Does the timing of reviews affect their precision?* In Section 7 we consider the relation of review timing and review precision. We did not detect strong dependencies between grading error and the time taken to complete a review, the order in which the student completed the reviews, or how late the reviews were completed with respect to the review deadline.

*Does error vary with class topic?* In Section 4 we consider the question of whether grading precision varies from topic to topic. Comparing broad topic areas, such as computer science, essays, science, we find the statistics to be quite similar, indicating how general factors are less important than the specifics of each class.

*Does tit-for-tat affect review feedback?* CrowdGrader allows students to leave feedback on the reviews and grades they receive; this feedback is then used as one of the factor that determines the student's grade in the assignment. The feedback was introduced to provide an incentive for writing helpful reviews. In Section 8 we show that when a grade is over 20% below the consensus, it receives a low feedback score due to tit-for-tat about 38% of the time.

In the next section, we give a brief description of CrowdGrader, and of the datasets on which our analysis is based. The subsequent sections present the details of the answers to the above questions. We conclude with a discussion on the nature of errors in peer grading, and on the implications for algorithms and reputation systems for computing consensus grades.

## 2. RELATED WORK

The accuracy of peer grading in the context of MOOCs has been analyzed in [12], where the match between instructor grade and student grades is analyzed in detail. The study finds a tendency by student to rate higher people that share their country of origin — and this in spite of the grading process being anonymous. The study finds that improvement in grading rubrics lead to improved grading accuracy. Geographical origin, along with gender, employment status, and other factors, are found to have influence on engagement in peer grading in a French MOOC in [4]. Our work is thus somewhat orthogonal to [4, 12]: we do not have data on student ethnicity, and we focus instead on factors measurable from the peer grading activity itself.

Frequently, peer grades are accompanied with reviewers' comments or feedback; [18] explores the possibility of using the review text to asses review quality. The authors show a successful application of classifiers and statistical Natural Language Processing to evaluate reviews.

*Peer Instruction* is a process in which students can observe grades by other reviewers, discuss the review, and consequently modify their grades [6]. The factors that influence grades in peer instruction have been studied in [2]. In spite of the different settings, [2] also observe that the behavior of high and low-scoring students is fairly similar in terms of their grading accuracy.

## 3. THE CROWDGRADER DATASET

To analyze the source of grading errors in peer grading, we rely on a dataset from CrowdGrader, a peer review and grading tool used in universities and high-schools [8]. After students submit their solutions to an assignment, students review and grade a certain number of submissions by their peers. From these peer grades, Crowdgrader computes a *consensus grade* for every submission. Once the review phase is concluded, the students can rate the reviews they received according to a 1 to 5-star rating. These review ratings are meant to provide an incentive for students to write detailed, helpful reviews of other students work.

The overall dataset we examined consisted in 288 assignments, for a total of 25,633 submissions and 113,169 reviews, written by 23,762 distinct reviewers. The number of reviewers is smaller than the number of submissions, as some students did not participate in the review phase. Table 1 gives a break-down of the dataset according to subject area. On average, each submission received 4.41 reviews, and each reviewer wrote on average 4.76 reviews.

We will refer to submissions also as *items*, and we will refer to students or reviewers also as *users,* thus adopting common terminology for general peer-review systems.

CrowdGrader includes three features that promote grading accuracy; these features likely influenced the data presented in this study.

*Incentives for accuracy.* The overall grade a student receives in a CrowdGrader assignment is a weighed average of the student's *submission, accuracy,* and *helpfulness* grades. The *accuracy grade* reflects the precision of the student's grade, compared either to the other grades for the same submission or, when available, to the instructor-assigned grade. The *helpfulness grade* grade reflects the rating received by the reviews written by the student. Combining the submission grade with the accuracy grage creates an incentive for students to be precise in their grading. The amount of incentive can be chosen by the instructor, but the default is to give 75% weight to the submission grade, 15% weight to the accuracy grade, and 10% weight to the helpfulness grade, and most instructors do not change this default.

*Ability to decline reviews.* Early in the development of CrowdGrader, we noticed that some of the most glaring grading errors occurred when reviewers were forced to enter a grade for submissions that they could not properly evaluate. This occurred, for in-

| | Assignments | Submissions | Reviewers | Reviews | Graded Assignments | Graded Submissions |
|---|---|---|---|---|---|---|
| Computer Science | 188 | 19397 | 17829 | 86347 | 68 | 2402 |
| Physics | 7 | 274 | 270 | 907 | 6 | 33 |
| Epidemiology | 5 | 337 | 313 | 1551 | 0 | 0 |
| Sociology | 49 | 3822 | 3683 | 18339 | 3 | 16 |
| Business | 26 | 1217 | 1108 | 3915 | 15 | 106 |
| English | 9 | 397 | 383 | 1717 | 1 | 7 |
| High-school | 7 | 279 | 278 | 1097 | 5 | 20 |
| Other | 4 | 189 | 176 | 393 | 0 | 0 |
| All Combined | 288 | 25633 | 23762 | 113169 | 93 | 2564 |

Table 1: The CrowdGrader dataset used in this study. *Graded assignments* are the assignments where an instructor or teaching assistant graded at least a subset of the submissions. *Graded submissions* is the number of submissions that were graded by instructors or teaching assistants, in addition to peer grading.

stance, when students could not open the files uploaded as part of the submission, due to software incompatibilities. To mitigate this problem, we gave students the ability to *decline* to perform reviews of particular submissions. The total number of submissions a student can decline is bounded, to prevent students from "shopping around" for the easiest submissions to review.

*Submission discussion forums.* Another early source of large errors in CrowdGrader consisted in gross mis-understandings between the author of a submission, and the reviewers. For instance, when zip archives are submitted, the reviewers may expect some information to be contained in one of the component files, whereas the author might have included it in another. Another example consists in mis-organizing the content of a software submission, so that the reviewers do not know how to run it and evaluate it. To remedy this, CrowdGrader introduced anonymous forums associated with each submission, where submission authors and reviewers can discuss any issues they encounter in evaluating the work.

# 4. ERRORS IN PEER GRADING

*Instructor grades and Olympic averages.* We measure review error as the difference between individual student grades, and the "consensus grade" for each submission. We consider two kinds of consensus grades. One is the *Olympic average* of the grades provided by the students: this is obtained by discarding the lowest and highest grade for each submission, and taking the average of the remaining grades. The other is the *instructor grade*. In CrowdGrader, instructors (or teaching assistants) have the option of re-grading submissions. In some assignments, instructors decided to grade most submissions as control; in other assignments, instructors mostly re-graded only submissions where student grades were in too much disagreement. When considering instructor grades, we consider only assignments of the first type, where instructors graded at least 30% of all submissions. Considering assignments where instructors grade only problematic submissions would considerably skew the statistics. The dataset, for instructor grades, is thus reduced to 19 assignments and 7675 reviews. Instructor and Olympic average grades have a coefficient of correlation $\rho = 0.81$ (with $p < 10^{-200}$), and an average absolute difference of 6.11 on the [0, 100] grading range.

*Global and per-topic errors.* Table 2 reports the size of errors in CrowdGrader peer grading assignments, split by assignment topic, and taking instructor grades and Olympic grades as reference. When the error is measured with respect to instructor grades, computer science, physics, and high-school assignments showed smaller average error than business, sociology and English, all of whose assignments required essay-writing. When the error is measured with respect to Olympic average, is is mainly business and English that show larger error.

| | Average Error | N. of Assignments |
|---|---|---|
| Computer Science | 7.52 | 15 |
| Physics | 10.6 | 1 |
| Business | 16.5 | 2 |
| English | 17.2 | 1 |
| High School | 10.6 | 1 |
| All | 7.67 | 19 |

(a) Error with respect to instructor grades, based on assignments with at least 30% of items graded by the instructor.

| | Average Error | N. of Assignments |
|---|---|---|
| Computer Science | 6.34 | 188 |
| Physics | 4.65 | 7 |
| Epidemiology | 4.57 | 5 |
| Sociology | 4.93 | 49 |
| Business | 7.7 | 26 |
| English | 8.37 | 9 |
| High School | 5.09 | 7 |
| Other | 8.15 | 4 |
| All | 6.16 | 288 |

(b) Error with respect to Olympic average.

Table 2: Mean absolute value difference error by topic. The grading range is normalized to [0, 100].

# 5. ITEM VS. STUDENT ERROR

We consider in this section the question of whether error can be attributed predominantly to imprecise students, or to items that are difficult to grade.

## 5.1 Average error behavior

To compare the contribution of students and items to grading errors, we develop a probabilistic model in which both students and items contribute to the evaluation error. The model is a modification of the $\mathbf{PG}_1$ model in [13], which allowed for student (but not item) error. In our model, each student has a *reliability* and each item has a *simplicity*; the variances of student and item errors are inversely proportional to their respective reliabilities and simplici-

| | Average Error | N. of Assignments |
|---|---|---|
| Computer Science | 19.6 | 15 |
| Physics | 14.4 | 1 |
| Business | 21.4 | 2 |
| English | 20.4 | 1 |
| High School | 14.4 | 1 |
| All | 19.6 | 19 |

(a) Error with respect to instructor grades, based on assignments with at least 30% of items graded by the instructor.

| | Average Error | N. of Assignments |
|---|---|---|
| Computer Science | 21.6 | 188 |
| Physics | 9.79 | 7 |
| Epidemiology | 9.38 | 5 |
| Sociology | 9.47 | 49 |
| Business | 11.2 | 26 |
| English | 20.1 | 9 |
| High School | 9.18 | 7 |
| Other | 11.5 | 4 |
| All | 19.6 | 288 |

(b) Error with respect to Olympic average.

Table 3: Root mean square error by topic. The grading range is normalized to [0, 100].

ties. Precisely:

$$\begin{aligned}
\text{(Reliability)}\ & \tau_u \sim \mathcal{G}(\alpha_0, \beta_0) \text{ for every student } u,\\
\text{(Simplicity)}\ & s_i \sim \mathcal{G}(\alpha_1, \beta_1) \text{ for every item } i,\\
\text{(True Grade)}\ & q_i \sim \mathcal{N}(\mu_0, 1/\gamma_0) \text{ for every item } i,\\
\text{(Observed Grade)}\ & g_{iu} \sim \mathcal{N}(q_i, 1/\tau_u + 1/s_i)\\
& \qquad \text{for every observed peer grade } g_{iu}
\end{aligned}$$

where $\mathcal{G}(\alpha, \beta)$ denotes the Gamma distribution with parameters $\alpha$, $\beta$, and $\mathcal{N}(q, v)$ denotes the normal distribution with average $q$ and variance $v$.

Given an assignment, we use Gibbs sampling [10] to infer the parameters $\alpha_0, \beta_0, \alpha_1, \beta_1, \mu_0, \gamma_0$. In order to apply Gibbs sampling, we need to start from suitable prior values for the quantities being estimated. To obtain suitable priors for the distribution of item quality, we first compute an estimated grade for each item using Olympic average, and we obtain $\mu_0$ and $\gamma_0$ by fitting a normal distribution to the estimated grades. To estimate prior parameters $\alpha_0, \beta_0$ of student reliabilities we fit a Gamma distribution to a set of approximated students reliabilities. In detail, for every student $u$ we populate a list of errors $l_u$ by the student. Again, we computer errors with respect to the average item grades after removing the extremes (the Olympic average). Using the list of error $l_u$, we estimate a standard deviation $\sigma_u$ for every student $u \in U$. This allows us to approximate student reliability $\hat{\tau}_u$ as $\frac{1}{\sigma_u^2}$. Prior parameters $\alpha_0, \beta_0$ are obtained by fitting a Gamma distribution to the set of estimated student reliabilities $\{\hat{\tau}_u | u \in U\}$. To estimate prior parameters $\alpha_1, \beta_1$ for item simplicities we use the same approach as for $\alpha_0, \beta_0$; the only difference is that item simplicities $\hat{s}_i$ are estimated using error lists $l_i$ computed for every item $i$, rather than for every student $u$.

Table 4 reports the average standard deviation of students and items inferred from the model. As we can see, students are responsible for over two thirds of the overall reviewing error.

| | students | items |
|---|---|---|
| Average Standard Deviation | 14.2 | 6.4 |

Table 4: The average standard deviation of students and items errors computed over 288 assignment with 25633 items. The grading range is [0, 100].

| | Error Threshold | | | | |
|---|---|---|---|---|---|
| | 10% | 15% | 20% | 25% | 30% |
| Students | 0.015 | 0.026 | 0.017 | 0.019 | 0.017 |
| Items | 0.075 | 0.082 | 0.082 | 0.1 | 0. 097 |

(a) Item errors computed with respect to instructor's grades. We use only assignments that have at least 30% of items grade by the instructor.

| | Error Threshold | | | | |
|---|---|---|---|---|---|
| | 10% | 15% | 20% | 25% | 30% |
| Students | 0.018 | 0.018 | 0.019 | 0.020 | 0.021 |
| Items | 0.045 | 0.030 | 0.020 | 0.021 | 0.020 |

(b) Item errors computed with respect to Olympic average.

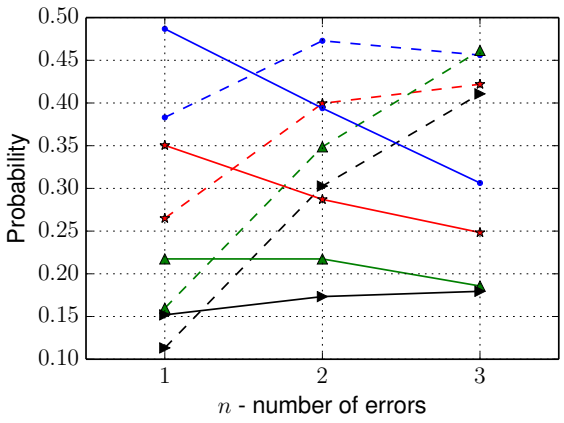
Table 5: Coefficient of constraint $I(X,Y)/H(X)$ of large errors on the same item or by the same student, for different error thresholds.

## 5.2 Large error behavior

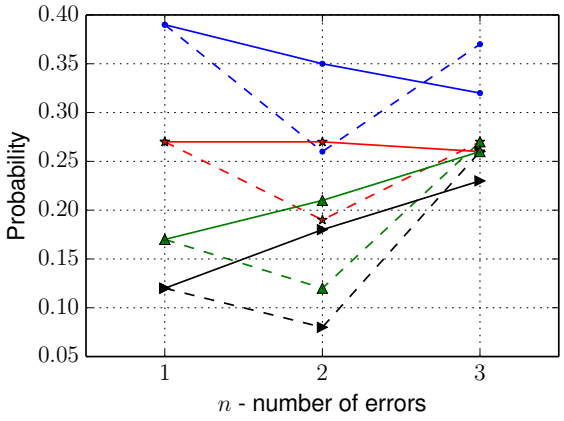
While students intuitively understand that small random errors will be averaged out, they are very concerned by large errors that, they fear, will skew their overall grade. Thus, we are interested in determining whether such large errors are more often due to students who are grossly imprecise, or items that are very hard to grade. In other words: do large errors cluster more around imprecise students, or around hard-to-grade items? We can answer this question because in CrowdGrader, items are assigned to students in a completely random way. Thus, any correlation between errors on items or students indicates causality.

We answer this question in two ways. First, we measured the information-theoretic *coefficient of constraint.* To compute it, let $X$ and $Y$ be two random variables, obtained by sampling uniformly at random two reviews $x$ and $y$ corresponding to the same item, or to the same student, and letting $X$ (resp. $Y$) be 1 if $x$ (resp. $y$) is incorrect by more than a pre-defined threshold (such as, 20% of the grading range for the assignment). Then, the mutual information $I(Y,X)$ indicates the amount of information shared by $X$ and $Y$, and the coefficient of constraint $I(X,Y)/H(X)$, where $H(X)$ is the entropy of $X$, is an information-theoretic measure of the correlation between $X$ and $Y$.

Tables 5 gives $I(X,Y)/H(X)$ for student and item errors, for different values of the error choice, and taking as reference truth for each item either the instructor grade, or the Olympic average for the item. When taking instructor grades as reference (Table 5a), large errors are about 5 times more correlated on items than on students, as measured by the coefficient of constraint. When Olympic grades are take as reference (Table 5b), large errors are about as correlated on items as they are on students. The difference in behavior is due to the fact that, when an instructor disagrees with the student-given grades on an item, this generates highly correlated errors on that item with respect to the instructor grade, but not with respect to the Olympic average. In any case, the results show that there is no particular correlation on students.




(a) Errors computed with respect to the instructor's grades. We use only assignments that have at least 30% of items grade by the instructor.




(b) Errors computed with respect to Olympic average.

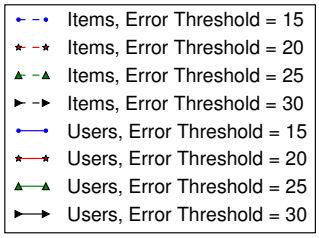



Figure 1: Conditional probabilities $\rho_n = P(\xi \geq n | \xi \geq n-1)$ of least $n$ errors given at least $n-1$ errors. We considered error thresholds of 15%, 20%, 25%, 30%.

Another way to measure whether large errors tend to cluster around hard-to-evaluate items or around imprecise students consists in measuring the conditional probability $\rho_n = P(\xi \geq n | \xi \geq n-1)$ of an item (resp. student) having $\xi \geq n$ grossly erroneous reviews, given than it has at least $n-1$. If errors on an item (resp. reviewer) are uncorrelated, we would expect that $\rho_1 = \rho_2 = \rho_3 = \cdots$. If these conditional probabilities grow with $n$, so that $\rho_3 > \rho_2 > \rho_1$, this indicates that the more errors an item (resp. a student) has participated in, the more likely it is that there are additional errors. The values of $\rho_1, \rho_2, \rho_3, \ldots$ allow thus one to form an intuitive appreciation for how clustered around items or students the errors are.

The results are given in Figure 1. The data shows some clustering around users, for large errors of over 30% of the grading range. However, clustering around users seems weaker than clustering around items.

This provides a possible explanation for why reputation systems have not proved effective in dealing with errors in peer-graded assignments with CrowdGrader. Reputation systems are effective in characterizing the precision of each student, and taking it into account when computing each item's grade. Our results indicate however that errors in CrowdGrader are not strongly correlated with students, limiting the potential of reputation systems.

# 6. STUDENT ABILITY VS. ACCURACY

A natural question is whether better students make better graders. To answer this question, we can approximate the expertise of every student with the grade received by the student's own submission, and we can then study the correlation between the student's submission grade, and the review error. As we have only partial coverage of students with instructor grades, we compute the grade received by the student's own submission via Olympic average, rather than instructor grade. As the two generally are close, this increases coverage with minimal influence on the results. We study grading error with respect to both instructor grades and Olympic average.

## 6.1 Aggregating data from multiple assignments

When aggregating data from multiple assignments, we cannot directly compare absolute values of grades, or absolute amount of time spent reviewing: each assignment has its own grade distribution, review time distribution, and so forth. To account for variation across assignments, we use the following approach. For each student there is an independent variable $x$, and an error $e$. In this section, $x$ is the grade received by the student's own submission, measured via Olympic average; in the next section, $x$ will be related to the time spent during the review, or the time at which the review is turned in. The error $e$ is the difference, for each review, between the grade assigned as part of the review, and the grade of the reviewed submission, obtained either via Olympic average or via instructor grading.

First, for each assignment independently, we sort all students according to their $x$-value, and we assign them to one of 10 percentile bins: if the assignment comprises $m$ students and the student ranks $k$-th, the student will be in the $\lceil 10k/m \rceil$ bin; we call these bins the 10%, 20%, ..., 100% bins. For each assignment $a$, we normalize the grading range to $[0, 100]$, and we let $n_{a,q}$ and $e_{a,q}$ be the number of students and the average error in the $q$ percentile bin of assignment $a$, respectively. The average error for assignment $a$ overall is thus $e_a = \sum_q n_{a,q} e_{a,q} / \sum_q n_{a,q}$. There are two ways of measuring the average error $e_{a,q}$ for one bin: as average absolute value error, or as average root-mean-square error. The two approaches lead to qualitatively similar conclusions, as we show later in this section.

We aggregate data from multiple assignments, computing for each percentile bin an absolute and a relative error, as follows. The *absolute* error $e_q$ for each percentile $q$ is computed as

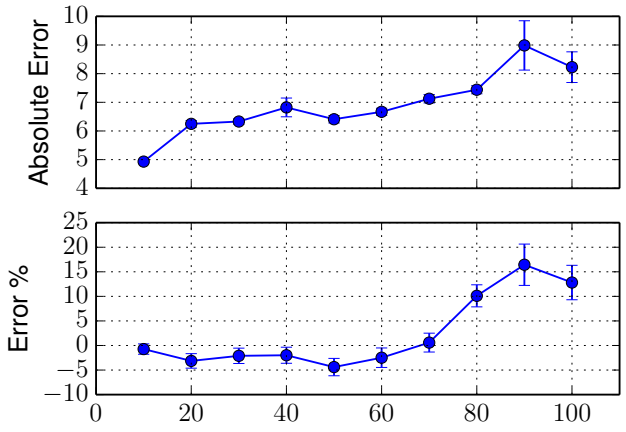
$$e_q = \sum_a n_{a,q} e_{a,q} \Big/ \sum_a n_{a,q}. \quad (1)$$

The *relative* error $r_q$ for each percentile $q$ is computed as

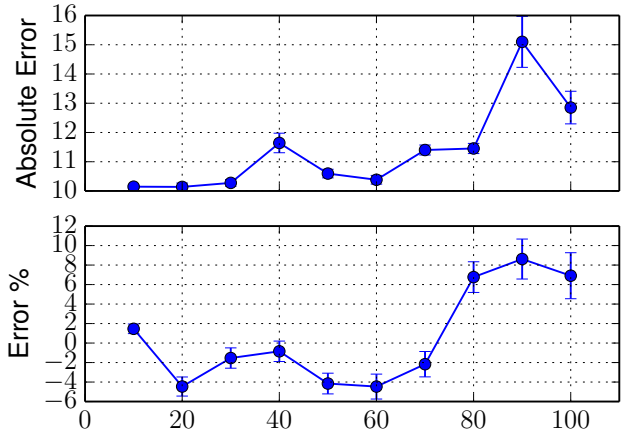
$$r_q = \sum_a n_{a,q} \big(e_{a,q}/e_a\big) \Big/ \sum_a n_{a,q}, \quad (2)$$

where $e_{a,q}/e_a$ is the relative error of bin $q$ in assignment $a$.

## 6.2 Student ability vs. error




(a) Mean absolute value difference error.

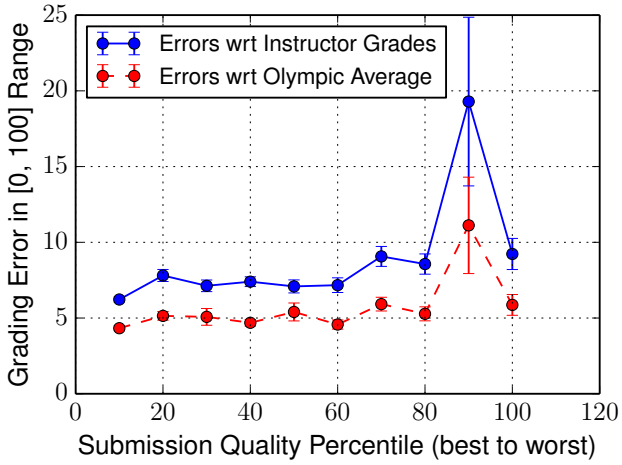



(b) Root mean square error.

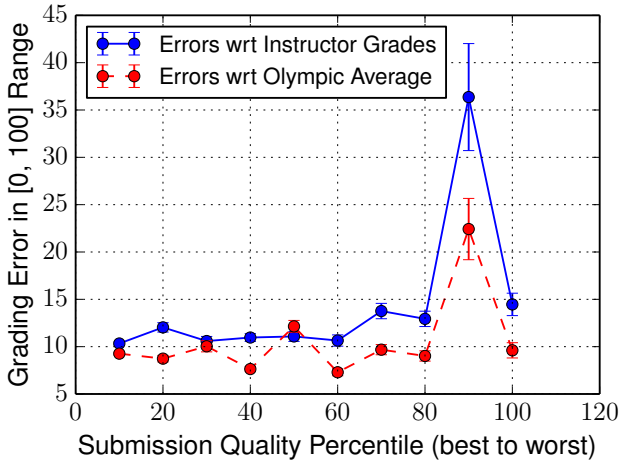
Figure 2: Average grading errors arranged into authors' submissions quality percentiles. Grading errors and submission qualities are measured with respect to the Olympic average grades. The first percentile bin 10% corresponds to reviewers that have authored submissions with highest grades. Error bars correspond to one standard deviation.

The data reported in Figure 2b shows the existence of some correlation between student submission grade, and grading precision, measured with respect to the Olympic average. In relative terms, students in the 80–100% percentile brackets show error that is 10% to 20% greater than students with higher submission grade. The absolute error tells a similar story. The two graphs do not have the same shape, due to the fact that relative errors are computed in (2) in a per-assignment fashion. In Figure 2a we report the same data, computed using rms error rather than average absolute value error. The data is qualitatively similar. Due to lack of space, in the remaining graphs we consider only average absolute error.

In Figure 3a we compare the error with respect to Olympic average with the error compared to instructor grades, for the subset of classes where at least 30% of submissions have been instructor-graded. While the absolute values are different, we see that the curves are very closely related, indicating that Olympic averages are a good proxy for instructor grades when studying relative changes in precision. The error with respect to instructor grades has very




(a) Mean absolute value difference error.




(b) Root mean square error.

Figure 3: Average grading error arranged into authors' submission quality percentiles. The first percentile bin 10% corresponds to reviewers that have authored submissions with highest grades. We report the error both with respect to instructor grades, and to the Olympic average, considering only assignments for which at least 30% of submissions have been graded by instructors. Error bars correspond to one standard deviation.

wide error bars for the 90% percentile, mainly due to the low number of data points we have for that percentile bracket in our dataset. We favor the comparison with the Olympic average, since the abundance of data makes the statistics more reliable.

The correlation between student ability (as measured by the submission score) and grading precision is lower than we expected. This might be a testament to the clarity of the rubrics and grading instructions provided by the instructors: apparently, such instructions ensure that most students are able to grade with reasonable precision the work by others. This may also be a consequence of the fundamental skill and background homogeneity of students in a classroom, as compared to a MOOC. We note that [2] also reported low correlation between student grades and student precision in the related setting of peer instruction.

## 7. REVIEW TIMING VS. ACCURACY

We next studied the effect of the time taken to perform the reviews, and the order in which they were performed, on review accuracy. These measurements are made possible by the fact that CrowdGrader assigns reviews one at a time: a student is assigned the next submission to review only once the previous review is completed. This dynamic assignment ensures that all submissions receive a sufficient number of reviews. If each student were pre-assigned a certain set of submissions to review, as is customary in conference paper reviewing, then students who omitted or forgot to perform reviews could cause some submissions to receive insufficient reviews. CrowdGrader records the time at which each submission is assigned for review to a student, and the time when the review is completed. For these results, to conserve space, we provide the error only with respect to the Olympic average, for which we have more data. A comparison of error with respect to Olympic average and instructor grades confirms that the Olympic average is a good proxy for studying variation with respect to instructor grade also We omit the analogous of Figure 3a.

*Time to complete a review.* We first considered the correlation between the time spent by students performing each review, and the accuracy of the review; the results are reported in Figure 4b. The results indicate that reviews that are performed moderately quickly tend to be slightly more precise. The correlation is weaker than we expected. We expected to find error peaks due to students that spent very little time reviewing, and that entered a quick guess for the submission grade, rather than performing a proper review. There are no such peaks: either students are very good at quickly estimating submission quality, or they mostly take reviewing and seriously in CrowdGrader. We believe the latter hypothesis is likely the correct one: for instance, in many computer science assignments, there is no good way of "eye-balling" the quality of a submission without compiling and running it.

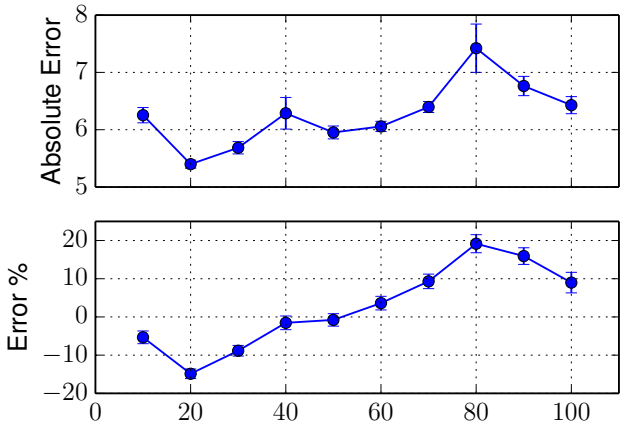
*Time at which a review is completed.* Next, we studied the correlation between the absolute time when reviews are performed, and the precision of the reviews. Figure 5a shows the existence of a modest correlation: the reviews that are completed in the first 10% percentile tend to be 10% more accurate than later reviews. The effect is rather small, however. In a typical CrowdGrader assignment, students are given ample time to complete their reviews, and the reviews themselves take only one hour or so to complete. Students likely do not feel they are under strong time pressure to complete the reviews, and time to deadline has little effect on accuracy.

*Order in which reviews are completed.* Lastly, we study whether the order in which a student performs the reviews affects the accuracy of the reviews. We are interested in the question of whether students learn while doing reviews, and become more precise, or whether they grow tired and impatient as they perform the reviews, and their accuracy decreases. Figure 6a shows that the accuracy of students does not vary significantly as the students progress in their review work. Evidently, the typical review load is sufficiently light that students do not suffer from decreased attention while completing the reviews.

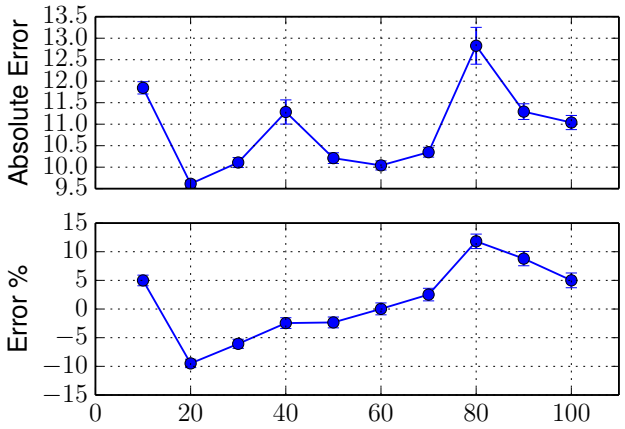
## 8. TIT-FOR-TAT IN REVIEW FEEDBACK

In CrowdGrader, students can leave feedback to each review and grade they receive. The feedback is expressed via 1-to-5 star rating systems as follows:

- 1 star: factually wrong; bogus.




(a) Mean absolute value difference error.



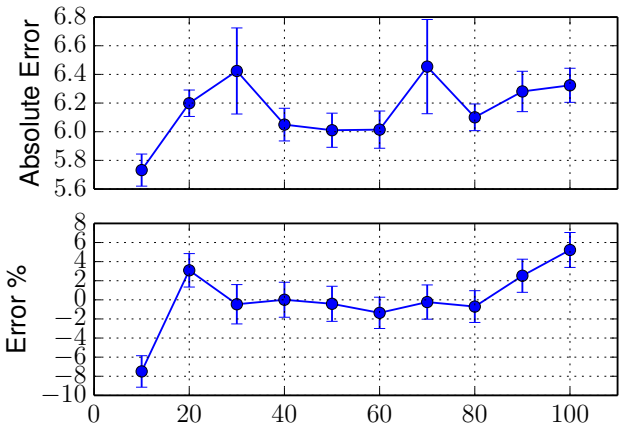

(b) Root mean square error.

Figure 4: Absolute and relative grading error vs. the time employed to perform a review; the first percentile bin 10% corresponds to reviews with shortest review time. The grading range is normalized to $[0, 100]$, and the error is measured with respect to the Olympic average. The error bars indicate one standard deviation.

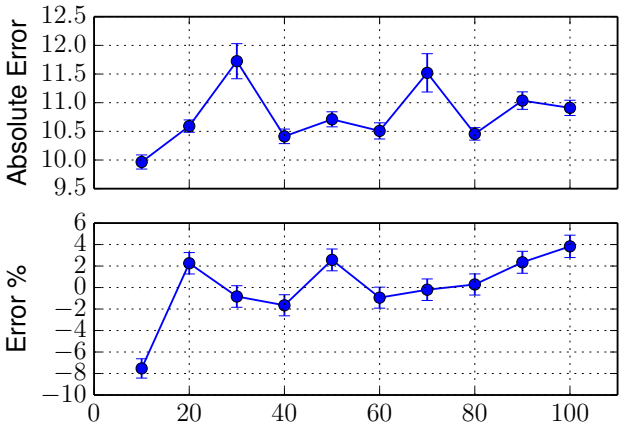
- 2 stars: unhelpful.
- 3 stars: neutral.
- 4 stars: somewhat helpful.
- 5 stars: very helpful.

Many such ratings are given as tit-for-tat: when a student receives a low grade, the student responds by assigning a low feedback score (typically, 1 star) to the corresponding review. Indeed, CrowdGrader includes a technique for identifying such tit-for-tat, so that students, whose overall grade depends also on the helpfulness of their reviews, are not unduly penalized. We were interested in analyzing the question of how prevalent tit-for-tat is.

Overall, review grade and review feedback have a correlation of $-0.39$, with a p-value smaller than $10^{-300}$. The negative correlation between grade and feedback indicates tit-for-tat, as there is no reason why lower grades should per-se be associated with written reviews that are less helpful. Interestingly, the negative correlation is fairly independent from the subject area. To bring the tit-for-tat

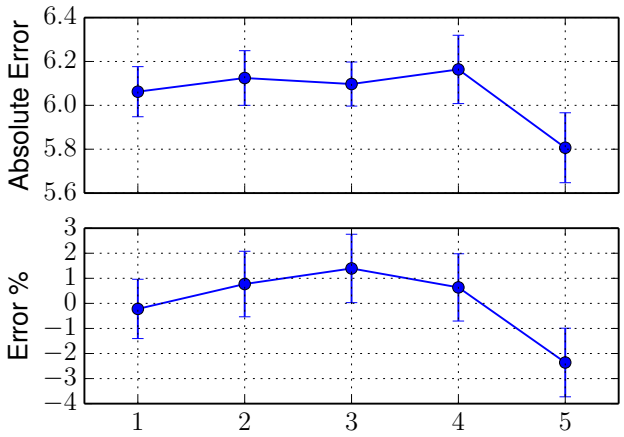



(a) Mean absolute value difference error.

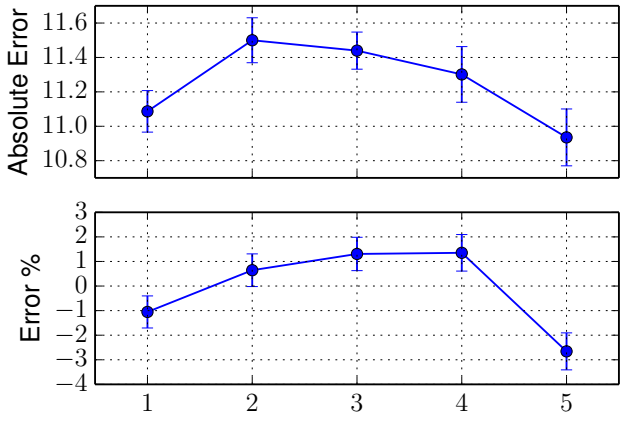



(b) Root mean square error.

Figure 5: Absolute and relative grading error vs. absolute time when a review is completed. The first percentile bin 10% corresponds to the 10% of reviews that were completed first among all assignment reviews. The grading range is normalized to $[0, 100]$, and the error is measured with respect to the Olympic average. The error bars indicate one standard deviation.




(a) Mean absolute value difference error.




(b) Root mean square error.

Figure 6: Absolute and relative grading error vs. ordinal number of a review by a student. The review 1 is the first a student performs, 2 is the second, and so forth. The grading range is normalized to $[0, 100]$, and the error is measured with respect to the Olympic average. Error bars indicate one standard deviation.

into sharper evidence, we computed also the following statistics. We consider a grade a $p$ (resp. $n$) outlier if the grade is over 20% above (resp. below) the Olympic average. We then measured the conditional probabilities $P_p$, $P_n$ that $p$ and $n$ outliers would receive a one or two-star rating, conditioned over the probability that the reviews received a rating at all (students do not always rate the reviews they receive). Over all assignments, we measured $P_p = 0.06$ and $P_n = 0.44$. Since there is no a-priori reason why overly negative reviews may be of worse quality than overly positive ones, the excess probability $P_n - P_p = 0.38$ can be explained by tit-for-tat. This shows that tit-for-tat is rather common: for grades that are 20% or more below the consensus, there is a 38% probability of low feedback due to tit for tat. Fortunately, it is easy to discard low ratings given in response to below-average grades, as CrowdGrader does.

## 9. DISCUSSION

We presented an analysis of a large body of peer-grading data, gathered on assignments that used CrowdGrader across a wide set of subjects, from engineering to business and humanities. Our main interest consisted in identifying the factors that influence grading errors, so that we could devise methods to control or compensate for such factors. Out results can be thus summarized:

- Large errors are no more strongly correlated on students than they are on items. In other words, students who are imprecise on many submissions are not a dominant source of error.
- There is some correlation between the quality of a student's own submission (which is an indication of the student's accomplishment), and the grading accuracy of the student, but the correlation is weak and limited to the student with highest, and lowest submission grades.
- There is little correlation between the accuracy of a review, and the time it took to perform the review, or how late in the review period the review was performed.
- There is clear evidence of tit-for-tat behavior when students give feedback on the reviews they receive.

All of the correlations we measured, except for the tit-for-tat one, are rather weak. This is a reassuring confirmation that peer-grading works as intended.There are no large sources of uncontrolled error due to factors such as student fatigue in doing the reviews, or gross inability of weaker students to perform the reviews. The peer-grading tool, in our classroom settings, ensures that the remaining errors are fairly randomly distributed, with little remaining structure.

The results highlight the difficulties in using reputation systems to compute submission grades in peer-grading assignments in high-school and university settings. Reputation systems characterize the behavior of each student, in terms for instance of their grading accuracy and bias, and compensate for each student's behavior when aggregating the individual review grades into a consensus grade. However, our results indicate that the large errors that most affect the fairness perception of peer grading are most closely associated with items, rather than with students. Reputation systems are powerless with respect to errors caused by hard-to-grade items: even if they can correctly pinpoint which submissions are hard to grade, little can be done except flagging them for instructor grading. Indeed, the reputation system approach of [13], which yielded error reductions of about 30% for MOOCs, yielded virtually no benefit in our classroom settings.

There is more potential, instead, in approaches that make it easier to grade difficult submissions. In CrowdGrader, we introduced anonymous forums, associated with each submission, where submissions authors and reviewers can discuss any issues that arise while reviewing the submission. These forums are routinely used, for instance, to solve the glitches that often arise when trying to compile or run code written by someone else. Anectodally, these forums have markedly increased the satisfaction with the peer-grading tool, as students feel that they have a safety net if they make small mistakes in formatting or submitting their work, and are in the loop should any issues occur.

## 10. REFERENCES


[1] S. P. Balfour. Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review (TM). *Research & Practice in Assessment*, 8, 2013.

[2] S. Bhatnagar, M. Desmarais, C. Whittaker, N. Lasry, M. Dugdale, and E. S. Charles. An analysis of peer-submitted and peer-reviewed answer rationales, in an asynchronous peer instruction based learning environment. *Proceedings of the 8th International Conference on Educational Data Mining*, 2015.

[3] D. Chinn. Peer assessment in the algorithms course. In *ACM SIGCSE Bulletin*, volume 37, pages 69–73. ACM, 2005.

[4] M. Cisel, R. Bachelet, and E. Bruillard. Peer assessment in the first french mooc: Analyzing assessors' behavior. *Proceedings of the 7th International Conference on Educational Data Mining*, 2014.

[5] S. Cooper and M. Sahami. Reflections on Stanford's MOOCs. *Communications of the ACM*, 56(2):28–30, 2013.

[6] C. H. Crouch and E. Mazur. Peer instruction: Ten years of experience and results. *American Journal of Physics*, 69(9):970–977, 2001.

[7] L. de Alfaro and M. Shavlovsky. Crowdgrader: Crowdsourcing the evaluation of homework assignments. *Technical Report UCSC-SOE-13-11, UC Santa Cruz, arXiv:1308.5273*, 2013.

[8] L. de Alfaro and M. Shavlovsky. CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proceedings of the 45th ACM technical symposium on Computer science education*, pages 415–420. ACM, 2014.

[9] E. F. Gehringer. Electronic peer review and peer grading in computer-science courses. *ACM SIGCSE Bulletin*, 33(1):139–143, 2001.

[10] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

[11] K. F. Hew and W. S. Cheung. Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. *Educational Research Review*, 12:45–58, 2014.

[12] C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer. Peer and self assessment in massive online classes. In *Design Thinking Research*, pages 131–168. Springer, 2015.

[13] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller. Tuned models of peer assessment in moocs. *arXiv preprint arXiv:1307.2579*, 2013.

[14] R. Robinson. Calibrated Peer Review: an application to increase student reading & writing skills. *The American Biology Teacher*, 63(7):474–480, 2001.

[15] J. Sadauskas, D. Tinapple, L. Olson, and R. Atkinson. CritViz: A Network Peer Critique Structure for Large Classrooms. In *EdMedia: World Conference on Educational Media and Technology*, volume 2013, pages 1437–1445, 2013.

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

[17] A. Venables and R. Summit. Enhancing scientific essay writing using peer assessment. *Innovations in Education and Teaching International*, 40(3):281–290, 2003.

[18] W. Xiong, D. J. Litman, and C. D. Schunn. Assessing reviewer's performance based on mining problem localization in peer-review data. In *EDM*, pages 211–220. ERIC, 2010.