Tuned Models of Peer Assessment in MOOCs

Chris Piech
Stanford University
piech@cs.stanford.edu

Jonathan Huang
Stanford University
jhuang11@stanford.edu

Zhenghao Chen
Coursera
zhenghao@coursera.org

Chuong Do
Coursera
cdo@coursera.org

Andrew Ng
Coursera
ng@coursera.org

Daphne Koller
Coursera
koller@coursera.org

ABSTRACT

In massive open-access online courses (MOOCs), peer grading serves as a critical tool for scaling the grading of complex, open-ended assignments to courses with tens or hundreds of thousands of students. But despite promising initial trials, it does not always deliver accurate results compared to human experts. In this paper, we develop algorithms for estimating and correcting for grader biases and reliabilities, showing significant improvement in peer grading accuracy on real data with 63,199 peer grades from Coursera's HCI course offerings — the largest peer grading networks analysed to date. We relate grader biases and reliabilities to other student factors such as engagement, performance as well as commenting style. We also show that our model can lead to more intelligent assignment of graders to gradees.

1. INTRODUCTION

The recent increase in popularity of massive open-access online courses (MOOCs), distributed on platforms such as Udacity, Coursera and EdX, has made it possible for anyone with an internet connection to enroll in free, university level courses. However while new web technologies allow for scalable ways to deliver video lecture content, implement social forums and track student progress in MOOCs, we remain limited in our ability to evaluate and give feedback for complex and often open-ended student assignments such as mathematical proofs, design problems and essays. Peer assessment — which has been historically used for logistical, pedagogical, metacognitive, and affective benefits [18] — offers a promising solution that can scale the grading of complex assignments in courses with tens or even hundreds of thousands of students.

Initial MOOC-scale peer grading experiments have shown promise. A recent offering of an online Human Computer Interaction (HCI) course demonstrated that on average, student grades in a MOOC exhibit some agreement with staff-given grades [13]. Despite their initial successes, there remains much room for improvement. It was estimated that 43% of student submissions in the HCI course were given a grade that fell over 10 percentage points from a corresponding staff grade, with some submissions up to 70pp from staff given grades. Thus a critical challenge lies in how to reliably obtain accurate grades from peers.

In this paper, we present the largest peer grading networks analysed to date with over $63,000$ peer grades. Our central contribution is to use this unprecedented volume of peer assessment data to extend the discourse on how to create an effective grading system. We formulate and evaluate intuitive probabilistic peer grading models for estimating submission grades as well as grader biases and reliabilities, allowing ourselves to compensate for grader idiosyncrasies. Our methods improve upon the accuracy of baseline peer grading systems that simply use the median of peer grades by over 30% in root mean squared error (RMSE).

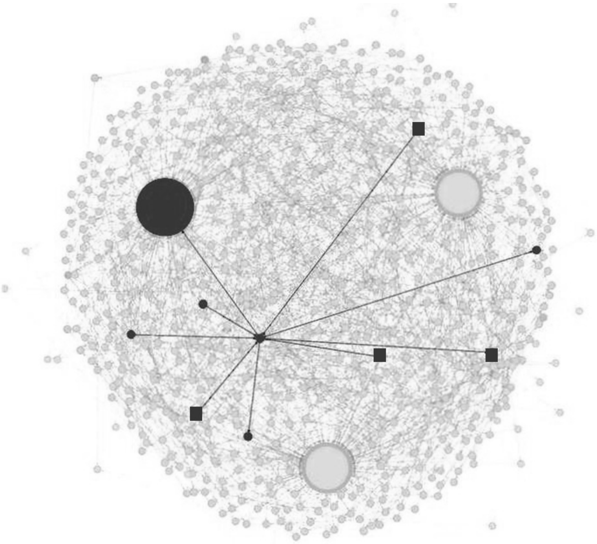


Figure 1: **Peer-grading network: Each node is a learner with edges depicting who graded whom. Node size represents the number of graders for that student. The highlighted learner shown above graded five students (circular nodes) and was in turn graded by four students (square nodes).**

In addition to achieving more accurate scoring for peer grading, we also show how fair scores (where our system arrives at a similar level of confidence about every student's grade) can be achieved by maintaining estimates of uncertainty of a submission's grade.

Finally we demonstrate that grader related quantities in our statistical model such as bias and reliability have much to say about other educationally relevant quantities. Specifically we explore summative influences: what variables correspond with a student being a better grader, and formative results: how peer grading affects future course participation. With the large amount of data available to us, we are

Table 1: Data Sets

	First HCI	Second HCI
Students	3,607	3,633
Assignments	5	5
Submissions	6,702	7,270
Peer Grades	31,067	32,132

able to perform detailed analyses of these relationships that would have been difficult to validate with smaller datasets. These results have notable relevance to the study of education in that the relationships observed depict hypotheses for the dynamics of students grading.

Because peer grading is structurally similar in both MOOCs and traditional brick and mortar classrooms, these results shed light on best practices across both mediums. At the same time, our work helps to describe the unique dynamics of peer assessment in a very new setting — one which may be part of a future with cheaper, more accessible education.

2. DATASETS

In this work, we use datasets collected from two consecutive Coursera offerings of Human Computer Interaction (HCI), taught by Stanford professor Scott Klemmer. The HCI courses used a calibrated peer grading system [17] in order to assess weekly student submissions for assignments which covered a number of different creative design tasks for building a web site. On every assignment, each student evaluated five randomly selected submissions (one of which was a "ground truth" submission, discussed below) based on a rubric[1], and in turn, was evaluated by four classmates. The final score given to a submission was determined as the median of the corresponding peer grades.[2] Peer grading was anonymized so that students could not see who they were evaluating, or who their evaluators were.

After the first offering (HCI1), the peer grading system was refined in several ways. Among other things, HCI2 featured a modified rubric that addressed some of the shortcomings of the original peer grading scheme and peer graders were divided into language groups (English and Spanish). Counting just those who submitted at least one assignment in the English offerings of the class, there were 3,607 students from the first offering (HCI 1) and 3,633 students from the second offering (HCI 2). These students came from diverse backgrounds (with a majority of students from outside of the United States). Collectively, these 7,240 students from around the world created 13,972 submissions, receiving 63,199 peer grades in total. See Table 1 for a summary of the dataset. To prevent overfitting our models to one instance of the class we used the data from HCI2 as a hold out set.

The software for the peer grading framework used by the HCI courses was designed to accommodate experimental validation of peer grading. A small number (3-5) of submissions for each assignment were set aside to serve as "ground truth." For each assignment, every student was given a ground truth submission to grade. Since there were thousands of students in the class, the ground truth submissions were "supergraded" (they had, on average, 160 assessments). Of note, the students were not told that one of the submissions they were assigned to mark belonged to the ground truth set. We hypothesized that by that law of large numbers, the mean of hundreds of student grades should tend towards the correct peer grade. In addition to having hundreds of student grades, the ground truth submissions were also evaluated by volunteer staff graders. However upon deeper analysis it seemed that the mean student grade was more consistently accurate with respect to the rubric than the volunteer staff grade which was not exempt from interrater inconsistency. In order to measure something as complex as a student's grade, it is more useful to take the average of hundreds of evaluations. Figure 1 shows the network of gradee-grader relationships on Assignment 5 of HCI1, where the three supergraded ground truth submissions are clearly visible.

3. PROBABILISTIC MODELS

The ideal peer grading system for a MOOC should satisfy the following desiderata: it should (1) provide highly reliable/accurate assessment, (2) allocate a balanced and limited workload across students and course staff, (3) be scalable to class sizes of tens or hundreds of thousands of students, and (4) apply broadly to a diverse collection of problem settings. In this section we discuss a number of ways to formulate a probabilistic model of peer grading to address these desiderata. The models that we introduce allow for us to algorithmically compensate for factors such as grader biases and reliabilities while maintaining estimates of uncertainty in a principled way. Through our paper, we will use v and u to refer to students in the peer grading network (each student is also a grader).

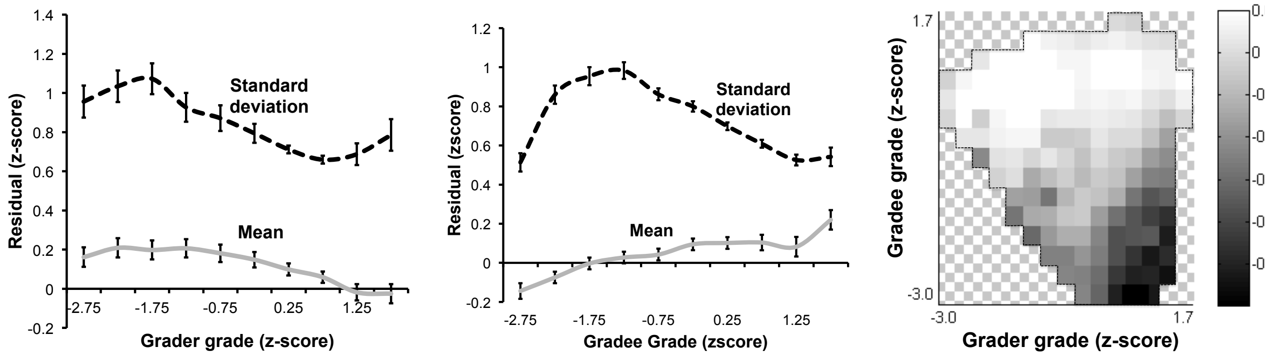
Our models assume the existence of the following quantities which are either observed or latent (unobserved) variables.

- **True scores:** We assume that every student u is associated with a submission that has a *true underlying score*, denoted s_u, which is unobserved and to be estimated.
- **Grader biases:** Every grader v is associated with a bias, $b_v \in \mathbb{R}$. These bias variables reflect a grader's tendency to either inflate or deflate her assessment by a certain number of percentage points.
- **Grader reliabilities:** We also model grader reliability, $\tau_v \in \mathbb{R}^+$, reflecting how close on average a grader's peer assessments tend to land near the corresponding submission's true score after having corrected for bias. Reliabilities will always correspond to the inverse variance of a normal distribution.
- **Observed grades:** Finally, $z_u^v \in \mathbb{R}$ is the observable score given by grader v to submission u.

Below we present, in order of increasing complexity, three statistical models that we have found to be particularly effective.

[1] See the Appendix: `http://stanford.edu/~cpiech/bio/papers/appendices/edm13_appendix.pdf`

[2] Our description is somewhat of a simplification — students also performed self-assessments and were given the higher of the median and their self grade provided that the two were within five percentage points of each other. We did not consider self assessments in this work.



(a)

(b)

(c)

Figure 2: (a) The relationship between a grader's homework performance (her grade) and statistics (mean/standard deviation) of grading performance (residual from true grade). (b) The relationship between a gradee's homework performance against statistics of assessments for her submissions. (c) Visualization of all three variables simultaneously, where intensity reflects the mean residual z-score. Empty boxes mean that there is not enough data available to compute a reliable estimate.

3.1 Model PG$_1$ (Grader Bias and Reliability)

Our first model, $\mathbf{PG}_1$ puts prior distributions over the latent variables and assumes for example that while an individual grader's bias may be nonzero, the average bias of many graders is zero. Specifically,

$$\begin{aligned}
&\text{(Reliability) } \tau_v \sim \mathcal{G}(\alpha_0,\, \beta_0) \text{ for every grader } v,\\
&\text{(Bias) } b_v \sim \mathcal{N}(0,\, 1/\eta_0) \text{ for every grader } v,\\
&\text{(True score) } s_u \sim \mathcal{N}(\mu_0,\, 1/\gamma_0) \text{ for every user } u, \text{ and}\\
&\text{(Observed score) } z_u^v \sim \mathcal{N}(s_u + b_v,\, 1/\tau_v),\\
&\qquad \text{for every observed peer grade.}
\end{aligned}$$

$\mathcal{G}$ refers to a gamma distribution. The variables α_0, β_0, η_0, μ_0 and γ_0 are hyperparameters for prior beliefs of each variable distribution. In our experiments, we also consider a simplified version of Model $\mathbf{PG}_1$ in which the reliability of every grader is fixed to be the same value. We refer to this simpler model in which only the grader biases are allowed to vary as $\mathbf{PG}_1$**-bias**.

3.2 Model PG$_2$ (Temporal Coherence)

The priors for the grader variables can play a particularly important role in the above model due to the fact that we typically only have about 4-5 grades to estimate the bias and reliability of each grader. A simple way to obtain more data per grader is to leverage observations made about the grader from previous assignments. To pose a model, we must understand the relationship of a grader's bias and reliability at homework T to that at homework T'. Is it the same or does it change over time?

To answer this question, we examine the correlation between the estimated biases from Model $\mathbf{PG}_1$ using the HCI1 dataset (see Section 2). Between consecutive assignments, a grader's biases have a Pearson correlation of 0.33 which represents a utilizable consistency. Grader reliability, on the other hand, has a low correlation. We therefore posit Model $\mathbf{PG}_2$ which allows for grader biases at homework T to depend on those at homework $T-1$ (and implicitly, on all prior homeworks). Specifically, Model $\mathbf{PG}_2$ assumes:

$$\begin{aligned}
&\tau_v^{(T)} \sim \mathcal{G}(\alpha_0, \beta_0) \text{ for every grader } v,\\
&b_v^{(T)} \sim \mathcal{N}(b_v^{(T-1)}, 1/\omega_0) \text{ for every grader } v,\\
&s_u^{(T)} \sim \mathcal{N}(\mu_0, 1/\gamma_0) \text{ for every user } u, \text{ and}\\
&z_u^{v,(T)} \sim \mathcal{N}(s_u^{(T)} + b_v^{(T)}, 1/\tau_v^{(T)}),\\
&\qquad \text{for every observed peer grade.}
\end{aligned}$$

Model $\mathbf{PG}_2$ requires that we normalize grades across different homework assignments to a consistent scale. In our experiments, for example, we have noticed that the set of grader biases had different variances on different homework assignments. Using a normalized score (z-score), however, allows us to propagate a student's underlying bias while remaining robust to assignment artifacts.

Note that while a model which captures the dynamics of true scores and reliabilities across assignments can be similarly imagined, we have focused only on the dynamics of bias for this work (which contributes the most towards improved accuracy while still being equitable).

3.3 Model PG$_3$ (Grader/Student Interplay)

A unique aspect of peer grading is that graders are themselves students with submissions being graded. Consequently, it is of interest to understand and model the relationship between one's grade and one's grading ability — for example, knowing that a student scored well on his assignment may be cause for placing more trust in that student as a grader.

In Figure 2, we show experiments exploring the relationships between the grader specific latent variables. In particular, we observe that high scoring students tend to be somewhat more reliable as graders (see details of the experiment in Section 4). Model $\mathbf{PG}_3$ formalizes this intuition by allowing the reliability of a grader to depend on her own grade:

$$\begin{aligned}
&b_v \sim \mathcal{N}(0,\, 1/\eta_0) \text{ for every grader } v,\\
&s_u \sim \mathcal{N}(\mu_0,\, 1/\gamma_0) \text{ for every user } u, \text{ and}\\
&z_u^v \sim \mathcal{N}\left(s_u + b_v,\, \frac{1}{\theta_1 s_v + \theta_0}\right),\\
&\qquad \text{for every observed peer grade.}
\end{aligned}$$

Note that Model $\mathbf{PG}_3$ extends $\mathbf{PG}_1$ by introducing new dependencies, allowing us to use a student's submission score to estimate her grading ability. At the same time Model

$\mathbf{PG}_3$ is more constrained, forcing grader reliability to depend on a single parameter instead of being allowed to vary arbitrarily, and thus prevents our model from overfitting.

Ethics and Incentives. If we are to use probabilistic inference to score students in a MOOC, the end goal could not simply be to optimize for accuracy. We must also consider fairness when it comes to deciding what variables to include in the model. It might be tempting, for example, to include variables such as race, ethnicity and gender into a model for better accuracy, but almost everyone would agree that these factors could not be fairly used within a scoring mechanism even if they improved prediction accuracy. Another example might be to model the temporal coherence of student grades (we observe a particularly strong temporal correlation between students' grades — with 0.46 Pearson coefficient — of consecutive homework assignments). But incorporating this temporal coherence for students scores into a scoring mechanism would not allow for students to be given a "clean slate" on each homework.

An interesting but subtle facet of $\mathbf{PG}_3$ is that by modelling a correlation between grader reliability and how well the grader did on the assignment, not only does getting a better grade on an assignment influence the model into thinking a particular student is a more reliable grader. It is also the case that if a student is a more reliable grader, the model is influenced into thinking that the student did better on the assignment. This relationship would allow for a student to "game" the mechanism into believing that they did better by grading as accurately as possible. Thus $\mathbf{PG}_3$ may in fact incentivize good grading. Giving students bonus points for better grading is not a new idea. However the nuance of $\mathbf{PG}_3$ is that these bonus points are justified in a statistical sense.

3.4 Inference and evaluation.

Given a probabilistic model of peer grading such as those discussed above, we would like to infer the values of the unobserved variables such as the true score of every submission, or the bias and reliability of each student as a grader. Inference can be framed as the problem of computing the posterior distribution over the latent variables conditioned on all observed peer grades.

Computing this posterior is nontrivial, since all of the variables are correlated with each other. For example, having good estimates of the biases of all of the graders to submission u would allow us to better estimate u's true score, s_u. However to estimate each bias b_v, we would have to have good estimates of the true scores of *all* of the submissions graded by v. We must therefore reason circularly.

To address this apparent chicken and egg problem, we turn to simple approximate inference methods. In the experiments reported in Section 4, we use Gibbs sampling [7], which produces a collection of samples from the (approximate) desired posterior distribution. These samples can then be used to estimate various quantities of interest. For example, given samples $s_u^1, s_u^2, \ldots, s_u^T$ from the posterior distribution over the true score of submission u, we estimate the true score as: $\hat{s}_u \equiv \frac{1}{T}\sum_{t=1}^{T} s_u^t$. We can also use the samples to quantify the uncertainty of our prediction by estimating the variance of the samples from the posterior, which we use in Section 4 when we examine peer grading efficiency. Note that while the ordinary Gibbs sampling algorithm can be performed in "closed form" for Models $\mathbf{PG}_1$ and $\mathbf{PG}_2$, Model $\mathbf{PG}_3$ requires numerical approximation due to the coupling of a submission's true score s_u with that of its grader, s_v. Visually we observe rapid mixing for our Gibbs chains, and in the experiments shown in Section 4, we use 800 iterations of Gibbs sampling, discarding the initial 80 burn-in samples.

Expectation-maximization (EM) is an alternative approximate inference approach, where we treat the true scores and grader biases as parameters and then use an iterative coordinate descent based algorithm to obtain point estimates of parameters. In practice, we find that both the Gibbs and EM approaches behave similarly. In general EM has the advantage of being significantly faster while obtaining posterior credible intervals is more natural using Gibbs. On the peer grading dataset the two methods produce analogous results. For example, $\mathbf{PG}_1$ with Gibbs and EM have RMSE scores of 5.42 and 5.43 on the first dataset respectively and with Gibbs running in roughly 5 minutes and EM running in 7 seconds. We refer the reader to the appendix for the full algorithmic details of Gibbs as well as EM.

Evaluation. To measure peer grading accuracy, we repeatedly simulate what score would have been assigned to each ground truth submission had it been peer graded. Our evaluation of how well we would have graded a single ground truth submission uses a two step methodology (based on the evaluation method of [13]): (1) We run inference using all of our data, except the peer grades of the ground truth submission being evaluated. This gives us an estimate of each grader's biases and reliabilities as well as model priors that were independent of the submission being evaluated. (2) We run simulations where we sampled four student assessments randomly from the pool of peer grades for the ground truth submission, estimate the submission's grade using the sample of assessments and record the residual between our estimated grade and the "true" grade. For each ground truth submission we run 3000 such simulations, from which we report the RMSE, the number of simulations which fell within five, and ten percentage points of the true score, the average standard deviation of the errors over each ground truth and the worst misgrade that the simulations produced.

We compare each of our probabilistic models to the grade estimation algorithm used on Coursera's platform. In the baseline model, the score given to students is the median of the four peer grades they received. Specifically, the baseline estimation does not take into account an individual grader's biases or reliabilities. Nor does it incorporate prior knowledge about the distribution of true grades.

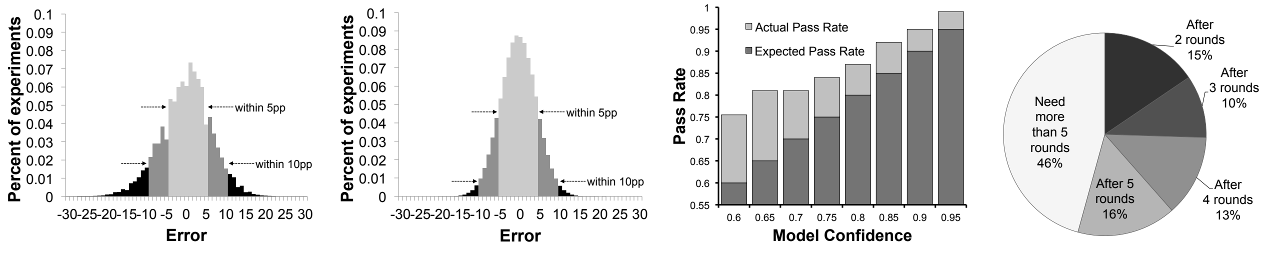
4. EXPERIMENTAL RESULTS

4.1 Accuracy of reweighted peer grading

Using probabilistic models leads to substantially higher grading accuracy. In our experiments we are able to reduce the RMS error on our prediction of the ground truth grade by 33% from 7.95 to 5.30. Similarly, on the second offering of the course we were able to reduce error by 31% from 6.43 to 4.73. For the second offering, this means that the num-

Table 2: Comparison of models on the two HCI courses

	HCI 1					HCI 2				
	Baseline	$\mathbf{PG_1}$-bias	$\mathbf{PG_1}$	$\mathbf{PG_2}$	$\mathbf{PG_3}$	Baseline	$\mathbf{PG_1}$-bias	$\mathbf{PG_1}$	$\mathbf{PG_2}$	$\mathbf{PG_3}$
RMSE	7.95	5.42	5.40	5.40	**5.30**	6.43	4.84	4.81	4.75	**4.73**
% Within 5pp	51	69	69	**71**	70	59	72	73	73	**74**
% Within 10pp	81	92	94	94	**95**	88	96	96	**97**	**97**
Mean Std	7.23	5.00	4.96	4.92	**4.77**	6.19	4.57	**4.52**	4.53	**4.52**
Worst Grade	-43	-34	**-30**	-32	**-30**	-36	-26	-26	**-25**	-26



(a) (b) (c) (d)

Figure 3: (a) Histogram of errors made using the baseline (median) scoring mechanism. (b) Histogram of errors using $\mathbf{PG_3}$. (c) A comparison of model confidence (x-axis) and actual success rate of predictions (y-axis), where being above the diagonal (dark bars) is better. (d) Number of submissions for which our model can declare "confidence" after K rounds of grading.

ber of students who received grades within 10 percentage points (pp) of their grade increased from 88% to 97%. Figures 3(a), 3(b) show the effect of using Model $\mathbf{PG_3}$ as a scoring mechanism on the histogram of grading errors and Table 2 shows the complete results for each model. Due to course improvements, we observe that students in HCI2 were significantly more consistent as graders compared to students in HCI1. However, we remark that every one of our models run on HCI1 outperforms the baseline grading system run on HCI2 with respect to every metric, indicating that the best gains in peer grading are likely to come from both an improved class design as well as statistical modeling.

Our results show that Models $\mathbf{PG_3}$ (with coupled grader score and reliability) and $\mathbf{PG_2}$ (with temporal coherence) yield the best results, with Model $\mathbf{PG_3}$ outperforming the other models with respect to most metrics. But the single change that provides the most significant gains in accuracy is obtained by estimating each grader's bias (Model $\mathbf{PG_1}$**-bias**). This simple model is responsible for 95% of our reduction in RMSE. The other changes all contribute comparatively smaller improvements to a more accurate model.

Our evaluation setup also allows us to test how accurate we would have been, had we had more than four grades per student. If the class had increased the number of grades that each student received to five (instead of four), our model could reduce RMSE error on the first and second offering of HCI to 4.19 and 4.36 respectively.

Surprisingly while modeling grader bias is particularly effective, modeling grader reliability does little to improve our performance. To dig deeper into this result we test our model on a synthetic dataset — one generated exactly from Model $\mathbf{PG_1}$. When using this synthetic data with only four grades per student it is difficult for the model to correctly estimate grader reliability. Modeling variance for each grader only seems to have a notable impact when students grade many assignments (more than 10). This experiment also suggests why $\mathbf{PG_3}$ is more useful than $\mathbf{PG_1}$. Though $\mathbf{PG_1}$ contains more expressive power than $\mathbf{PG_3}$, estimating only two parameters for grader reliability (θ_0 and θ_1) is more statistically tractable with only four grades per student than estimating a reliability, τ_v, for each grader.

4.2 Fairness and efficiency in peer grading

One of the advantages of using a probabilistic model for peer grading is that we can obtain a belief distribution over grades (as opposed to a single score) for each student. These distributions give us a natural way of calculating how confident the model is when it predicts a grade for a student which opens up the possibility of a more equitable allocation of graders. For example, at a given point midway through the peer grading process, our model may be highly confident in its prediction for a given student's score, but very unsure in its prediction for another student. In this situation, to ensure that each student gets fair access to quality feedback, we could reassign graders to gradees such that submissions which have low-confidence scores are given to more and/or better graders.

The first step towards more fair allocation of grades is to ask ourselves: how accurate are our estimates of confidence? For example, we would like to know how to interpret what it means in practice when our Bayesian model is 90% confident that its prediction of a learner's true score is within 10pp of the actual true score.

To better understand our confidence estimates, we run the following experiment: We first performed a large number of peer grading simulations on ground truth. From each simulation we calculate how confident our model is that the grade it predicts is within 5%, 7%, and 10%, of the true score, respectively. We then bin the estimated confidences into ranges 0-5%, 5-10%, etc. After collecting over 5000 predictions per range, we test the pass rate of each range.

For example, suppose we select four assessments of the same ground truth submission in a simulation. If our model reports a 72% confidence — based on those four assessments — that our predicted grade is within 5pp of the true score, we add that estimate to the set of predictions in the 70% to 75% confidence range. When we test this confidence range the example prediction "passes" if its estimate is in fact within 5pp of the ground truth score.

One worry is that our model might be overconfident about its predictions even when wrong. However the results, shown in Figure 3(c), demonstrate that our confidence estimates are on the conservative side — for example over 95% of the time that our model claims it is between 90 and 95% confident of a prediction, the model's estimate is correct.

Since we have reason to believe that our confidence values are accurate, we can employ our posterior belief distributions to better allocate grades. To understand how much benefit we could get out of improved grade allocation, we estimate at what point in the grading process we were confident about each submission's score. For each homework assignment, we simulate grading taking place in rounds. In the first round, we only include the first grade submitted by each grader (which may have been a ground truth grade). In the second round, we included the first two, etc. For each round we run our model using the corresponding subset of grades and count the number of submissions for which we are over 90% confident that our predicted grades were within 10pp of the student's true grade.

After only two rounds of grading we are highly confident in our estimated grade for 15% of submissions (this generally means that the submission has a grade close to the assignment mean, and has two similar grades from graders). Figure 3(d) shows how the set of confident submissions grows over the grading rounds. Our experiment demonstrates a clear opportunity for grades to be reallocated as well as a pressing need for some submissions to get more grades. For 54% of students, after all rounds, we are still unsure of their submission's true score.

4.3 Graders in the context of the MOOC

Applying probabilistic models to peer grading networks allows us to increase our grade accuracy and better allocate what submissions students should grade. Another product of our work is an assignment — with a belief distribution — for a true score, grader bias and grader reliability for each student. We can use this large dataset to derive new understanding about peer grading as both a formative and summative assessment. We focus our investigation on two questions, (1) what factors influence how well a student grades? and (2) how does grading ability affect future class performance in a MOOC?

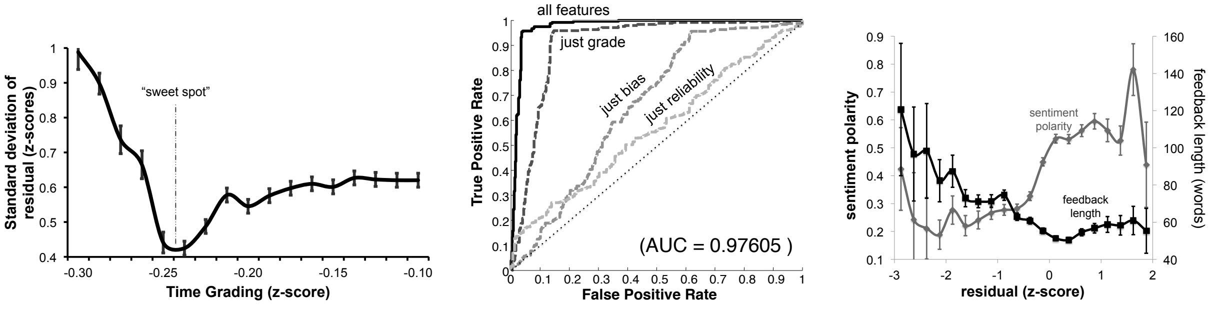
Influential factors for grader ability. To explore what factors influence how well a student grades we compare grading residual (how far off a grader's score is from our model estimated true score) to: time spent grading, grader grade, and gradee grade.

Time spent grading shows a particularly interesting trend (Figure 4(a)). As hypothesized, students that "snap grade" their peers' work (the students whose time spent grading has a z-score of less than -0.30), are both unreliable (the variance of their residuals is over 1 standard deviation away from the gradee's true score) and tend to slightly inflate grades. More surprising is that over the tens of thousands of grades, there is a "sweet spot" of time spent grading. Students who grade assessments with a time that has a z-score of around -0.25 have significantly lower residual standard deviations (with p-value < 0.001, diff $= 0.3$ standard deviations) than students who take a long time to grade (i.e., time spent grading has a z-score $>$ -0.20). This sweet spot is only visible when we look at normalized grading times. For most assignments in the HCI class, the sweet spot corresponds to around 20 minutes grading. This may reflect both that with any less time a grader does not have enough of a chance to fully examine her gradee's work, and that a long grading session may mean that the grader had trouble understanding some facet of the submission.

Examining the relationship between grader grade, gradee grade and how they affect the residual also shows a set of notable trends. Graders that score higher on assignments have close to monotonically decreasing biases (Figure 2(a)). Getting a better grade on the homework in general makes students more reliable graders; with the notable exception that the students that get the best grades (+1.75 z-score) are not as accurate as the students who do very well (+.75 z-score, $p = 0.04$). The superlative submissions — both the best and the worst — are the easiest to grade, and the submissions which are one standard deviation below the mean are the hardest (Figure 2(b)). Finally, our results show that students are least biased when grading peers with similar score (Figure 2(c)). The best students significantly downgrade the worst submissions and the worst students notably inflate the best submissions.

In addition to numerical scores, graders were asked to provide feedback in the form of free form text comments to their gradees. In order to understand the relationship between grading performance and commenting style, we compare grading residual against the comment length as well as sentiment polarity of the comment (Figure 4(c)). To measure the polarity of a comment, we use the sentiment analysis word list from [15] and implement a simple sentiment analyzer that returns a (normalized) polarity score (positive or negative) proportional to the sum of word valences over the comment. For both comment length and polarity, we filter out all non-English words. We observe that comments that correspond to larger negative residuals are typically significantly longer, suggesting perhaps that students write more about the weaknesses of a submission than strong points. That being said, we observe that overall, the comments mostly range in polarity from neutral to quite positive, suggesting that rather than being highly negative to some submissions, many students make an effort to be balanced in their comments to peers.

Grader ability and future performance. We also tested what signal grading ability has with predicting future participation. Based on the theory that the best graders are intrinsically motivated, we hypothesized that being a reliable grader would add a different dimension of information to a student's engagement which we should be able to use to



(a) (b) (c)

Figure 4: **(a) Grader consistency (measured using standard deviation of grading residual) as a function of time spent grading. (b) ROC curve comparing performance (with linear SVM) at predicting future class participation given a student's grade, bias, reliability or all three. (c) Commenting style (length of comment and sentiment polarity) as a function of grading residual.**

better predict future engagement. We tested this hypothesis by constructing a classification task in which we predict whether a student would participate in the next assignment (or conversely which students would "drop out"). In addition to the student's grade, we experimented with including grader bias and reliability as features in a linear classifier. Our results (Figure 4(b)) show that including grader bias and reliability improved our predictive ability by 5pp from an area under the curve (AUC) score of 0.93 to an AUC of 0.98. Properties about how a student grades, captures a dimension of their engagement which is missed by their assignment grade.

5. RELATED WORK

The statistical models we present in this paper are part of a long tradition of models which have been proposed for the purposes of aggregating information from noisy human labelers or workers. Many of these works adapt classical item-response theory (IRT) models [3] to the problem of "grading without an answer key" and appear in the literature from educational aptitude testing [10, 16, 14], to cultural anthropology [4, 12], and more recently to HCI in the context of human computation and crowdsourcing [19]. In educational testing, for example, Johnson [10] and Rogers et al. [16] propose models for combining human judgements of essays. These papers analyze *dedicated human graders* who each evaluated hundreds of essays, allowing for a rich model to be fitted on a per-grader basis. In contrast, with peer grading in MOOCs, each student only assesses a handful of assignments, necessitating more constrained models.

In a recent paper, and in a setting perhaps most similar to our own, Goldin et al. [9, 1, 8] use Bayesian models for peer grading in a smaller scale classroom setting. As in our own work, [8] posits a grader bias, and in fact incorporates rubric-specific biases, but does not consider many of the issues raised here such as grading task reallocation or the relationship between grader bias and student engagement, for example.

One of the central themes of the crowdsourcing literature, that of balancing label accuracy against labor cost, is one which MOOC peer grading systems must contend with as well. In such problems, one typically receives a number of noisy labels (for example in an image tagging task) and the challenge lies in (1) resolving the "correct" label (often discrete, but sometimes continuous) and (2) deciding whether to hire more labelers for a given task. Explosion of interest in recent years has led to widespread applications of crowdsourcing [2, 11]. For example in image annotation, Whitehill et al. [19] present a method similar to our own in which they model discrete "true image labels" as well as labeler accuracy. While our work draws from the crowdsourcing literature, the problem of peer grading is unique in several ways. For example, the fact that the graders are also gradees in peer grading is quite different from typical crowdsourcing settings in which there is a dichotomy between the labelers and the items being labeled, and motivates different models (such as Model $\mathbf{PG}_3$). In crowdsourcing applications, the end goal often lies in determining the true labels rather than to understand anything about the labelers themselves, whereas in peer grading, as we have shown, the insights that we can glean about the graders have educational value.

A similar problem to peer-grading is the paper assignment problem for the peer review process in academic conferences. While related in that the central challenge of both problems involves fusing disparate human opinions about open-ended creative work, many of the specific challenges are distinct. For one, side information plays a much larger role in peer review, where conference chairs typically rely heavily on personal or elicited knowledge of reviewer expertise or citation link structure to assign reviewer roles [5]. Peer grading on the other hand seems less sensitive to personal preferences, where a single submission should be equally well graded by a large fraction of students in the course.

6. DISCUSSION AND FUTURE WORK

Our paper presents methods for making large scale peer grading systems more dependable, accurate, and efficient. We show that there is much to be gained by maintaining estimates of grader specific quantities such as bias and reliability. In addition to improving peer grading accuracy by up to 30%, these quantities give unique insights into peer grading as a formative and summative assessment. Due to these promising results, our implementation is being used and evaluated in the third offering of HCI.

Moreover, trends among the latent variables that are coloured in by our model suggest results that could benefit the understanding of peer grading as a pedagogical tool. For example, understanding the factors that cause the "sweet spot" in grading time that we observed would be helpful in teaching graders. The trend between time spent grading and reliability also raises the question of how to incentivize students to spend enough time grading to provide careful and high quality feedback to their peers, particularly in an openacess course. Using model $\mathbf{PG}_3$ for scoring, as we discussed, makes a student's score dependent on grading performance, and may be one way to build a justified incentive directly into the scoring mechanism. There are also a number of challenging theoretical open questions on the mechanism design issues behind peer grading which [6] has taken steps to address. Similarly, the subtle patterns between grader score and reliability that are visible given our volume of data add an interesting piece of evidence to explore from an educational perspective. While, it is unsurprising that good students grade better, we also observe that the very best students in the class, were worse graders than students in the 70th percentile. What is the force behind this trend?

There remain a number of issues to be addressed in future work. We have considered the problem of determining which submissions need to be allocated additional graders. However, deciding which grader is best for evaluating a particular submission is an open problem whose solution could depend on a number of variables, from the writing styles of the grader and gradee to their respective cultural or linguistic backgrounds, a particularly important issue for the global scale course rosters that arise in MOOCS. Moreover, in our study we ovserved that the mean of hundreds of students who graded the same assignment was more reliable than volunteer staff grades. This points to a more in-depth investigation into how accurate "expert" grades really are. Finally, it is not clear how to present scores which are calculated by a complicated peer grading model to students. While this communication might be easy when a student's final grade is simply set to be the mean or median of peer grades, does each student need to know the inner workings of a more sophisticated statistical backend? Students may be unhappy with the lack of transparency in grading mechanisms, or on the other hand might feel more satisfied with their overall grade.

As MOOCs become more widespread, the need for reliable grading and feedback for open ended assignments becomes ever more critical. By addressing the shortcomings of current peer grading systems, we hope that students everywhere can get more from peer grading and consequently, more from their free online, open access educational experience.

Acknowledgments

We thank Chinmay Kulkarni and Scott Klemmer for providing assistance with the HCI datasets and Leonidas Guibas and John Mitchell for discussions and support. Jonathan Huang is supported by an NSF Computing Innovations Fellowship.

7. REFERENCES

[1] K. Ashley and I. Goldin. Toward ai-enhanced computer-supported peer review in legal education. In *24th International Conference on Legal Knowledge and Information Systems (JURIX)*, volume 235, 2011.

[2] Y. Bachrach, T. Minka, J. Guiver, and T. Graepel. How To Grade a Test Without Knowing the Answers - A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing. In *The 29th Annual International Conference on Machine Learning*, ICML '12, 2012.

[3] F. Baker. *The basics of item response theory.* ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, 2001.

[4] W. H. Batchelder and A. K. Romney. Test theory without an answer key. *Psychometrika*, 53:71–92, 1988.

[5] L. Charlin, R. Zemel, and C. Boutilier. A framework for optimizing paper matching. In *Proceedings of Uncertainty in Artificial Intelligence (UAI'11)*, 2011.

[6] A. Dasgupta and A. Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22th international conference on World wide web. ACM, 2013.*, 2013.

[7] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE PAMI*, (6):721–741, 1984.

[8] I. Goldin. Accounting for peer reviewer bias with bayesian models. In *Proceedings of the Workshop on Intelligent Support for Learning Groups at the 11th International Conference on Intelligent Tutoring Systems*, 2012.

[9] I. M. Goldin and K. D. Ashley. Peering inside peer review with bayesian models. In *Proceedings of the 15th international conference on Artificial intelligence in education*, AIED'11, pages 90–97, Berlin, Heidelberg, 2011. Springer-Verlag.

[10] V. E. Johnson. On bayesian analysis of multi-rater ordinal data: An application to automated essay grading. *Journal of the American Statistical Association*, 91:42–51, 1996.

[11] E. Kamar, S. Hacker, and E. Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *In AAMAS*, 2012.

[12] G. Karabatsos and W. Batchelder. Markov chain estimation for test theory without an answer key. *Psychometrika*, 68(3):373–389, 2003.

[13] C. Kulkarni, K. Pang-Wei, H. Le, D. Chia, K. Papadopoulos, D. Koller, and S. R. Klemmer. Scaling self and peer assessment to the global design classroom. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013.

[14] R. J. Mislevy, R. G. Almond, D. Yan, and L. S. Steinberg. Bayes nets in educational assessment: Where the numbers come from. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence*, pages 437–446. Morgan Kaufmann Publishers Inc., 1999.

[15] F. Å. Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[16] S. Rogers, M. Girolami, and T. Polajnar. Semi-parametric analysis of multi-rater data. *Statistics and Computing*, 20(3):317–334, July 2010.

[17] A. A. Russell. Calibrated peer review - a writing and critical-thinking instructional tool. *Teaching Tips: Innovations in Undergraduate Science Instruction*, page 54, 2004.

[18] P. M. Sadler and E. Good. The impact of self-and peer-grading on student learning. *Educational assessment*, 11(1):1–31, 2006.

[19] J. Whitehill, P. Ruvolo, T. fan Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, pages 2035–2043. MIT Press, 2009.