Practical Peer Prediction for Peer Assessment

Victor Shnayder
Harvard SEAS; edX
33 Oxford St., Cambridge MA
shnayder@seas.harvard.edu

David C. Parkes
Harvard SEAS
33 Oxford St., Cambridge MA
parkes@eecs.harvard.edu

Abstract

We provide an empirical analysis of peer prediction mechanisms, which reward participants for information in settings when there is no ground truth against which to score reports. We simulate the mechanisms on a dataset of three million peer assessments from the edX MOOC platform. We evaluate different mechanisms on score variability, which is connected to fairness, risk aversion, and participant learning. We also assess the magnitude of the incentives to invest effort, and study the effect of participant coordination on low-information signals. We find that the *correlated agreement mechanism* has lower variation in reward than other mechanisms. A concern is that the gain from exerting effort is relatively low across all mechanisms, due to frequent disagreement between peers. Our conclusions are relevant for crowdsourcing in education as well as other domains.

1 Introduction

We study the crowdsourcing of information in applications where it is difficult or expensive to verify contributions. There are many possible settings, including reporting information about businesses to improve products such as Google Maps, assessing peer work in large-scale education, and eliciting emotional reactions to video content or images. The objective is to encourage individuals to invest effort and make reports that reflect their viewpoint, even when this may be a minority viewpoint. Given reports, algorithmic methods can be used to aggregate the information in different ways.

The paradigm of *peer prediction* adopts explicit rewards to promote effort and truthful reports. In the absence of gold standard answers and the ability to verify reports, these rewards are determined based on comparisons between reports from different participants. Peer prediction has been studied for more than a decade, and there are now a number of mechanisms that have attractive theoretical properties—needing minimal information to operate, having broad domains of applicability, placing low reporting burden on participants, and avoiding undesirable "group-think" style equilibria.

As far as we know, peer prediction has not yet been deployed in any large-scale application.[1] Peer assessment in MOOCs is an exciting application for peer prediction—done well, it would enable low-cost and thus broadly accessible education in subjects that are difficult to automatically assess, such as writing, design, and public speaking.

Previous work on peer prediction has focused on the design of mechanisms that are *proper* (truthfulness is an equilibrium) and *strong truthful* (truthfulness is the equilibrium with highest score) (Shnayder et al. 2016). We study several previously unexplored mechanism properties that matter for practical deployment. First, the magnitude of the benefit of exerting effort and being truthful over uninformed strategies is important: in educational settings, scaling the scores arbitrarily to increase the relative value of effort is impossible within a fixed grade range. Second, participants may be risk averse, and prefer more certain strategies, even with lower expected scores. Last, it is important to strive for fairness: participants who evaluate their peers equally well should be rewarded equally. These concerns apply in education and other applications.

As a step toward deployment, we evaluate four candidate peer prediction mechanisms on a dataset of three million peer assessments from the edX MOOC platform. The comparison mechanisms include the classic output agreement mechanism as well as more recent designs (Kamble et al. 2015; Radanovic, Faltings, and Jurca 2016; Shnayder et al. 2016). Our analysis is not experimental—we take existing peer assessments from a system that does not evaluate scorers, and compute what peer prediction mechanisms would do given these reports. Our key results:

- The benefit from exerting effort in evaluating peers is relatively low across all candidate mechanisms, due to relatively low agreement between peer scores.
- The *correlated agreement mechanism* (Shnayder et al. 2016) has lower reward variation than other candidate mechanisms, because it rewards reports even without exact agreement between peers.

[1] There is some empirical work on peer prediction: Gao et al. (2014) study equilibrium selection in a simple binary setting, showing that agents find collusive equilibria. In contrast, Faltings, Pu, and Tran (2014) study a many-signal setting where uninformed equilibria did not appear to be a problem. We are not aware of any systematic studies of peer prediction in massive open online courses (MOOCs), though Radanovic, Faltings, and Jurca (2016) present some initial positive experimental results from an on-campus experiment.

- The low peer agreement in our data set makes all mechanisms susceptible to student coordination on easy-to-see but unintended signals, as described by Gao, Wright, and Leyton-Brown (2016).

In all cases, increasing agreement between peers would make peer prediction more practical. This can be accomplished by rubric design and student training, as well as using peer prediction itself to encourage effort. Because our data comes from a system without incentives for accurate reporting, our paper should be read as suggesting new evaluation criteria for peer prediction mechanisms, and raising a question about the necessary levels of agreement between peers, to be answered by future studies of deployed mechanisms. We hope that the low peer agreement we report does not discourage such experiments.

2 Peer Assessment in Education

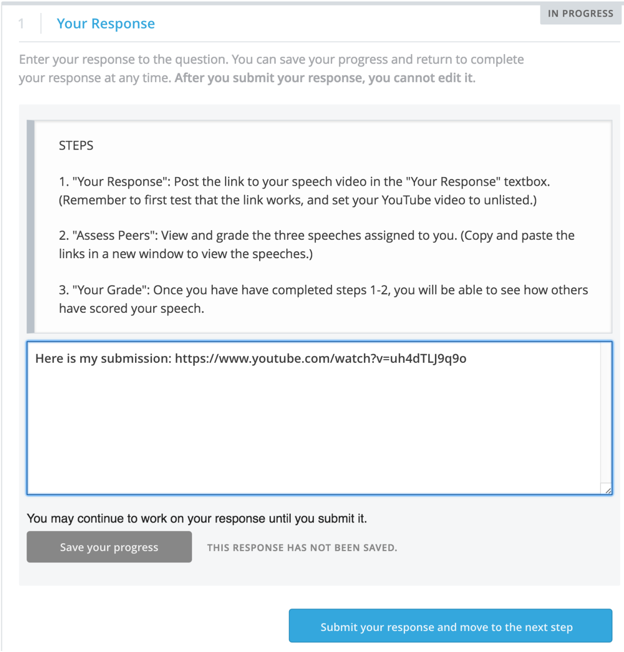
Peer assessment has a long history in education (see e.g. Goldfinch and Raeside (1990) and Falchikov (1995)) and is part of the much broader field of peer learning, which includes many types of peer-to-peer interaction in formal and informal settings.

For readers unfamiliar with peer assessment, we briefly summarize some lessons from its use in the classroom, to give a broader context for our incentive-focused study. There are two primary concerns about the scores given in peer assessments. The first is *reliability*, whether peers agree with each other. If not, ratings will have high variance, and many graders will be needed for each assignment to get a good estimate. The second is *validity*, whether the average peer score is "right" (Cho, Schunn, and Wilson 2006). In the typical situation where there is no absolute notion of right, it is typical to compare with instructor grades.

Calibrated peer review (Russell 2004) helps improve validity and reliability: before students assess each other, they practice grading three instructor-created samples of varying quality until they give the right grades. Good rubric design is also critical to reliability. Orsmond and Merry (1996) note that objective evaluation criteria are easier to assess, especially if the rater does not need to be an expert in the subject to distinguish between the possibilities.

In the last several years, peer assessment has been deployed in massive online courses at much larger scales than before. As a concrete example, Figure 1 shows a screenshot of the edX peer assessment system. Students submit their responses, and are paired randomly for review.

Research in large scale peer assessment has focused primarily on evaluating students' skill at assessment and compensating for grader bias (Piech et al. 2013), as well as helping students self-adjust for bias and provide better feedback (Kulkarni et al. 2013). Piech et al. (2013) test several models of student bias and reliability, testing for temporal coherence in bias as well as correlation between high scoring and being more reliable as a grader. Kulkarni et al. (2013) compare peer and staff grading, and find that the median peer grades are quite close to staff grades.



Enter your response to the question. You can save your progress and return to complete your response at any time. **After you submit your response, you cannot edit it.**

STEPS

1. "Your Response": Post the link to your speech video in the "Your Response" textbox. (Remember to first test that the link works, and set your YouTube video to unlisted.)

2. "Assess Peers": View and grade the three speeches assigned to you. (Copy and paste the links in a new window to view the speeches.)

3. "Your Grade": Once you have have completed steps 1-2, you will be able to see how others have scored your speech.

Here is my submission: https://www.youtube.com/watch?v=uh4dTLJ9q9o

You may continue to work on your response until you submit it.

THIS RESPONSE HAS NOT BEEN SAVED.

Figure 1: Screenshot from the edX peer assessment system, for a public speaking assignment.

Other recent studies focus on other aspects of peer assessment. PeerStudio (Kulkarni, Bernstein, and Klemmer 2015) improves learning by ensuring fast feedback in large scale peer assessment. The *Mechanical TA* (Wright and Leyton-Brown 2015) study focuses on reducing TA workload in high-stakes peer grading by reducing the need to spot-check peer grades.

Outside of peer assessment, many behavioral economics studies have shown that expected reward is not all-important in determining how people behave in practice (see Erev and Roth (2014) for a review). In our study, we are particularly motivated by *risk aversion*, which causes people to choose lower expected reward for more consistency, with high payoff variability leading to more random choices (Erev and Barron 2005; Busemeyer and Townsend 1993).

3 Peer Prediction Mechanisms

Peer prediction mechanisms are modelled as follows: agents are assigned to tasks, and observe a *signal* for each task that encodes the information the system wants to elicit. The signal model includes a *signal prior* $P(s)$, the probability that an agent observes signal s, as well as a *signal* joint $P(s, s')$, the probability that two agents who do the same task get signals s and s', respectively.

Agents report their signals, either truthfully as observed, or strategically to increase their expected score or avoid the effort of observing the signal precisely in the first place. Some mechanisms also require reporting information beyond the observed signal.

The mechanism compares reports, and computes a reward for each report. A basic goal is for the mechanism to be (strictly) *proper*, so that truthful reporting is a (strict) correlated equilibrium—if other agents are truthful, being truthful oneself is (strictly) a best response given the shared tasks. We restrict our study to *minimal* mechanisms, which do not require any additional information beyond a signal report, as these are more practical. We only include *detail-free* mechanisms, where the reward computation does not depend on precise details of the probabilistic signal model.[2]

We compare the following mechanisms:

Output Agreement (OA) (von Ahn and Dabbish 2004). For each report r, the system picks a reference report r'—another agent's report on the same task—and defines score $\sigma(r, r') = A(r, r')$, where $A(x, y)$ is the agreement function, defined to be 1 if $x = y$, and 0 otherwise. The OA mechanism is only strictly proper when the signal distribution is *self-dominant*, meaning that a user's observation is also the most likely observation for their reference peers.

We include OA in our study because of its simplicity. However, it and other early peer prediction mechanisms allow agents to coordinate and get higher rewards by reporting untruthfully. In OA, all agents simply reporting the same thing each time guarantees maximal reward.[3]

The next two mechanisms use the empirically observed report prior and joint distributions. We denote these $\hat{P}(\cdot)$ and $\hat{P}(\cdot, \cdot)$, respectively.

Robust Peer Truth Serum (RPTS) (Radanovic, Faltings, and Jurca 2016). This is a version of OA in which scores are scaled based on observed report frequencies; the system collects all reports, computes the empirical prior $\hat{P}(r)$ of each report r, and defines score $\sigma(r, r') = A(r, r')/\hat{P}(r)$, where r' is a reference report, just as in OA. This results in higher scores for matches on uncommon reports, which has two benefits: the mechanism requires a weaker *self-predicting* condition on the signal model—seeing a signal should increase the likelihood peer agents observe the same signal—and constant reporting now has lower expected score than truthfulness. RPTS requires that the number of tasks is large enough to make the empirical prior accurate.

Kamble (Kamble et al. 2015). This is another scaled version of OA. The system collects all reports, computes the empirical joint $\hat{P}(r, r')$, and defines score $\sigma(r, r') = A(r, r')/\sqrt{\hat{P}(r, r)}$, or 0 if $\hat{P}(r, r)$ is exactly 0 or 1. Similarly to RPTS, constant reporting again has lower expected score than truthfulness. Additionally, the mechanism is proper for general signal distributions.

Correlated Agreement (CA) (Shnayder et al. 2016). The CA mechanism is *multi-task*, so each agent reports on several tasks (at least two). An agent is rewarded for being more likely to match reports of peers doing the same task than the reports of peers doing other tasks. Let r_i^k denote the report received from agent i on task k. The mechanism is described, w.l.o.g., for two agents, 1 and 2:

1. Assign three or more tasks to the agents, two or more tasks per agent, including at least one overlapping task. Let M_s, M_1, and M_2 denote the shared, agent-1 and agent-2 tasks, respectively.
2. The score for a shared task $k \in M_s$ to each agent is
$$\sigma_k = \Lambda(r_1^k, r_2^k) - \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \Lambda(i, j) \cdot h_{1,i} \cdot h_{2,j}, \qquad (1)$$
where $\Lambda : \{0, \dots, n-1\} \times \{0, \dots, n-1\} \to \mathbb{R}$ is a score matrix with $\Lambda(s, s') = 1$ if $P(s, s') > P(s)P(s')$, and 0 otherwise, $h_{1,i} = \frac{|\{\ell \in M_1 | r_1^\ell = i\}|}{|M_1|}$ is the empirical frequency with which agent 1 reports signal i in tasks in set M_1, and $h_{2,j} = \frac{|\{\ell \in M_2 | r_2^\ell = j\}|}{|M_2|}$ is the empirical frequency with which agent 2 reports signal j in tasks in set M_2. This definition for Λ rewards agreement on positively correlated pairs of signals.
3. The total score to an agent is the sum of the score across all shared tasks.

CA is proper (not strictly), and *informed truthful*: the payoff for all agents being truthful is weakly higher than any other strategy profile, and strictly higher than any uninformed (signal-independent) reporting strategy (Shnayder et al. 2016). CA works with small numbers of tasks if the designer knows the direction of correlation between pairs of signals, needed to define Λ, or can learn these correlations from agent reports when there are many tasks.

3.1 Scaling Scores

To be practical for peer assessment, a mechanism's scores must be positive, and have bounded range—like any other grade, course teams need a way to say that assessing peers on an assignment counts for a particular number of points, and it is unreasonable to tell students that they may get an unboundedly high score with a very small probability, compensating for much more likely low scores.[4]

We set the scores for all mechanisms to be in $[0, 1]$ to make comparisons consistent. For RPTS and the Kamble mechanism, we do this by "clamping"—imposing a minimum on the report prior $\hat{P}(r)$ and the joint factor $\sqrt{\hat{P}(r, r)}$, respectively, and scaling to ensure the resulting score is in $[0, 1]$. We choose the minimum value to balance between effective score range and frequency of clamping in our dataset—whenever clamping applies, it breaks the mechanism's theoretical guarantees, effectively underpaying for

[2] We omit the scoring-rule based mechanism of Miller, Resnick, and Zeckhauser (2005), because it is not detail-free and not strong truthful, and non-minimal mechanisms (Prelec 2004; Witkowski and Parkes 2012a; 2012b; Radanovic and Faltings 2013). We also omit the minimal, strong truthful mechanism in Radanovic and Faltings (2015), because it requires many more reports per task than are typical in peer assessment.

[3] Jurca and Faltings (2009) attempted to fix this by rewarding near-agreement, not perfect agreement, with several peers. Dasgupta and Ghosh (2013) went further to design the first mechanism that guaranteed *strong truthfulness*, where the truthful equilibrium has higher payoff than all other equilibria, in settings with binary reports. Peer assessment uses non-binary reports, so we study several newer mechanisms that provide similar guarantees with arbitrary numbers of signals.

[4] The bounded range means that the standard theoretical trick of linearly scaling payoffs until the difference between truthful reporting and other strategies is big enough is not viable.

Category	Example	Count
Courses	*"Eating, Then and Now"*	254
Submission prompts	*"What is food?"*	682
Evaluation criteria	*"Correct grammar"*	1983
Submissions	*"Cheese is the best food"*	354312
Peer assessments	*3/5 points*	3090452

Table 1: Dataset summary, with examples of each item. There are approximately 1500 assessments for an average evaluation criterion.

unlikely reports.[5] An undesirable side-effect of this adjustment is that typical reports, with high priors by definition, will only use a small fraction of the score range, and only unlikely reports with prior close to the minimum will get scores close to 1.

For the CA mechanism, we remapped scores from the base range of $[-1, 1]$ into $[0, 1]$. The effect is that the expected score for uninformed reporting is 0.5, regardless of the reports of other learners.

All the mechanisms use a single reference peer as described, and can be modified to give the average score over several such peers. For example, for OA, given a set of reference reports $r_1, \ldots, r_n$ from n different peers on the same task, the mechanism could instead give score

$$\sigma(r) = \frac{1}{n} \sum_{i=1}^{n} A(r, r_i), \tag{2}$$

and similarly for the other mechanisms. Choosing a random reference peer or averaging across all reference peers has no effect on the expected score of a mechanism, but does affect the score variability. We study both variants below in Section 7 (see Figure 11).

4 The edX Dataset

The dataset in our study consists of peer assessments from edX, a site that offers open online courses, and includes data from 2014-2016.[6] Each peer assessment is a tuple

```
(course, item, submitter, submission,
submission_time, scorer, criterion,
points),
```

corresponding to a scorer assessing the given submission along a particular evaluation criterion, and giving it a score.

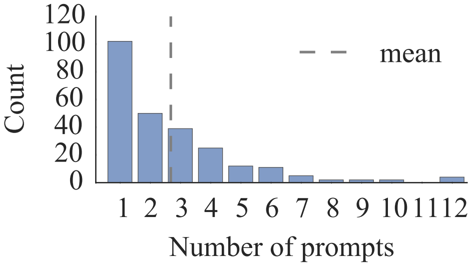


Figure 2: Prompts per course.

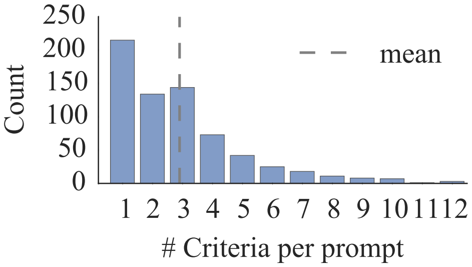


Figure 3: Histogram of evaluation criteria per prompt.

As a preprocessing pass, we keep only the latest evaluation for each `(scorer, submission)` pair, and discard criteria with fewer than 100 assessments. This leaves just over three million assessments, across about 2000 evaluation criteria, in 254 courses. Table 1 shows summary counts.

Submissions for each prompt are assessed on several evaluation criteria. For example, a short essay in a writing class may be judged on four criteria: grammar, style, argument, and appropriate citations. Figure 3 shows the number of evaluation criteria per prompt. Most prompts have four or fewer evaluation criteria. For each evaluation criterion, students can select a point value corresponding to a particular rubric option (e.g. 5/5, "Perfect grammar."), and each criterion induces a separate, empirical signal distribution. Most courses in our dataset only used peer assessment one or two times; a few courses had weekly or bi-weekly peer assessment assignments. Figure 2 shows the full distribution of the number of prompts per course.

Many evaluation criteria have several hundred assessments (median 733), with a few from large courses having ten thousand or more (Figure 4). The mean is about 1500.

4.1 Probabilistic Models for Reports

We now explore the details of the assessments for different prompts, looking at the number of options (i.e., the number of possible *signals*) for different evaluation criteria, the probabilities of those signals, and the correlation structures between them.

Figure 5 shows the distribution of the number of distinct options per criterion. An initially surprising observa-

[5] For RPTS, we make the minimal prior value 0.1, and divide scores by 10. This makes the expected score for uninformed reporting is 0.1. For Kamble, we make the minimum value of $\sqrt{\hat{P}(r, r)}$ 0.25 and divide scores by 4. These values balance between clamping too often and having the typical scores use a significant fraction of the $[0, 1]$ range.

[6] A summarized dataset of the joint report probabilities for each of the 1983 evaluation criteria is available at `https://github.com/HarvardEconCS/shnayder-peer-prediction-analysis`. The full dataset of individual students' assessments is sensitive, and cannot be shared.

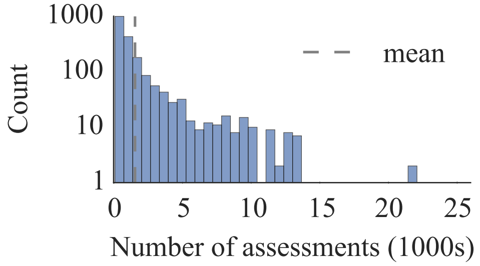


Figure 4: Histogram of the number of assessments across evaluation criteria. Note the log scale.

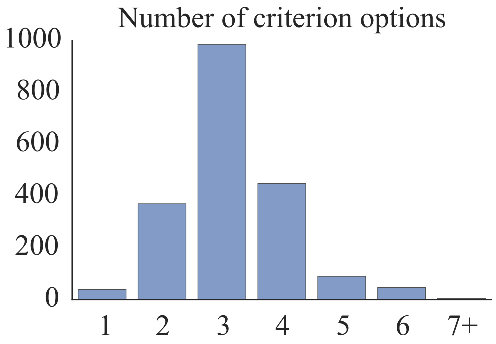


Figure 5: Histogram of models by number of criterion options (possible signal values).

tion: there are several criteria where students had one option in "assessing" their peer. The explanation is creative course teams using the peer assessment tool for open ended peer feedback, without wanting numerical assessment. We ignore these criteria in our analysis. On the other extreme, there was a course team that specified 21 score options for a criterion. Going forward, we focus on just models with two to six score options, since they account for the vast majority of the data.

Our dataset contains student reports, not the true signals observed by students. This is the best we can do without a prohibitive amount of manual grading, and we assume that the reports are a noisy approximation of the true signals. Since students are participating in a free class without much outside incentive for completion, doing the evaluation at all is indicative of exerting some effort.[7] From here on, we use *report* and *signal* interchangeably, unless explicitly distinguished.

We call a given signal distribution, corresponding to an evaluation criterion in the dataset, a *model*. Let $P(s)$ represent the prior probability of an agent seeing signal s on an arbitrary task. Let $P(s, s')$ denote the joint probability that two agents will see signals s and s'. We are also interested in what the joint distribution would be if signals were independent but with the same prior. This is the product-of-marginals distribution, written $Q(s, s') = P(s)P(s')$. Finally, we use $P(s'|s)$ to denote the signal posterior: the probability an agent observes s', conditioned on another agent observing s on the same task.

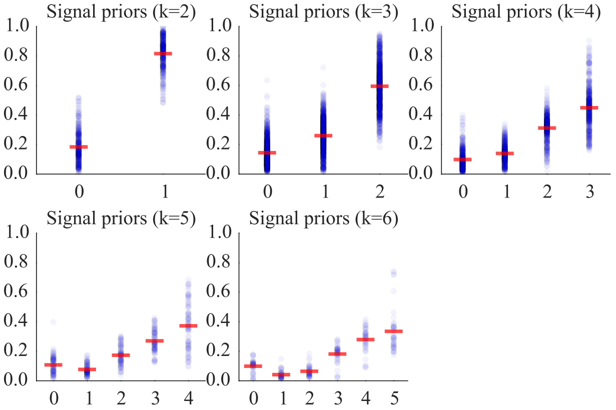


Figure 6: Signal priors by number of possible signal values, with all distributions for fixed number of signals on the same axis. The red line marker is the average for that signal. There is significant variation among models, with a clear trend toward higher scores.

We look next at the signal priors $P(s)$, which are important to the design, applicability, and robustness of peer-prediction mechanisms. The prior for an evaluation criterion is the probability with which each score appears. Figure 6 shows a plot for each number of signals k, plotting all distributions of size k on one plot, along with the average values. There is significant variation, but the priors are clearly non-uniform, with higher values more likely. An interesting secondary feature is that for $k \in \{5, 6\}$, non-zero scores below the median (1, and $\{1, 2\}$, respectively) tend to go unused. This suggests that most submissions are either very bad or incomplete, or ok-to-great, with few in between.[8]

An obvious question about a peer assessment system is whether peers usually agree. We look at this in several ways. A summary metric is the probability that two random peers assessing the same submission will report the same assessment. This probability is 61% in our dataset. To give a baseline, the probability that two peers assessing random submissions to the same prompt would agree is 52%. The high baseline probability makes sense in light of the non-uniform priors. Since many reports are of the highest possible signal (61% overall), two random assessments for different submissions frequently agree on that by chance. Figure 7 shows the distributions of observed agreement of same-submission reports and "default" agreement, if two reports for different submissions are chosen for a particular criterion. The vertical lines give the mean probabilities across the dataset, with

[7] Nevertheless, as MOOCs start to provide credentials based on peer-assessed work, we believe it will become increasingly important to provide explicit credit mechanisms for peer assessment.

[8] It also suggests that course teams may be able to simplify their rubrics, giving fewer options without losing many meaningful distinctions.

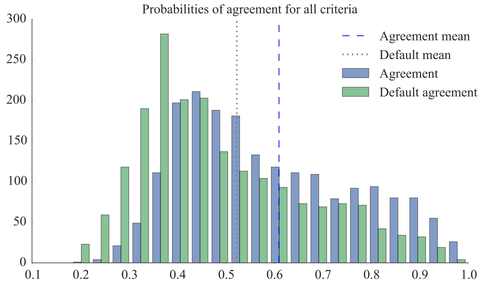


Figure 7: Histograms of observed and submission-independent “default” agreement between reports, per evaluation criterion. The means are weighted by the number of assessments for that criterion.

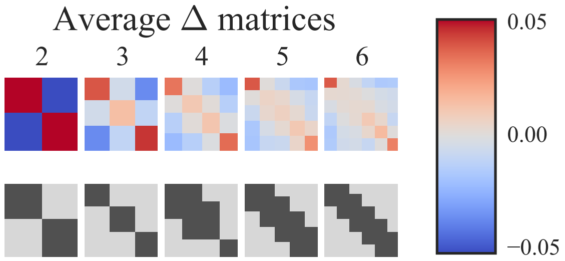


Figure 8: Average delta matrices and their sign structure. The positive areas along the diagonal correspond to the ordinal structure of peer assessment—nearby signals are likely to be positively correlated.

criteria weighted by number of reports.

Another way to look at agreement is to look at the correlation between pairs of signals. For this, define the *Delta matrix* Δ (Shnayder et al. 2016), an $n \times n$ matrix, with entry (i, j) defined as

$$\Delta_{s,s'} = P(s, s') - P(s)P(s'), \tag{3}$$

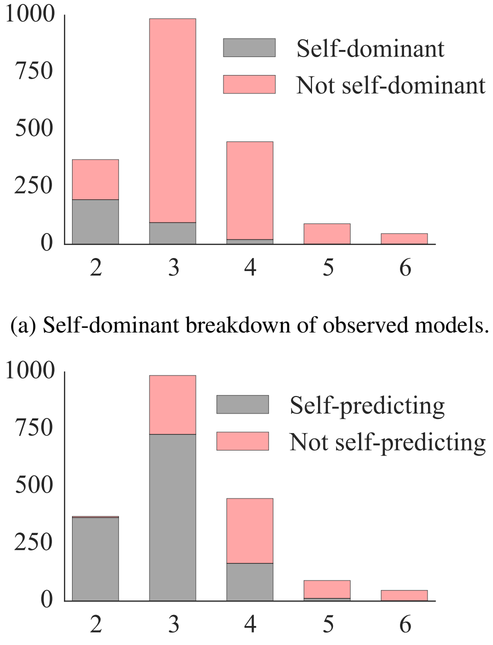
or equivalently as the difference between the joint and product-of-marginals distributions:

$$\Delta = P(\cdot, \cdot) - Q(\cdot, \cdot) \tag{4}$$

The delta matrix encodes the correlation (positive or negative) between different realized signals. The average values in this Delta matrix, grouped by prompts with the same number of reports, are shown in Figure 8, along with the sign structure which gives the direction of the correlation. As expected in a setting where the signal values are ordered,[9] the correlations are positive along the diagonal—if one student thinks the right score is 3/5, it increases the likelihood that their peer will say 2, 3, or 4.

5 Analysis I: Appropriateness

As a first analysis step, we look at how often the technical conditions required for the validity of different peer prediction mechanisms hold in our dataset.



(a) Self-dominant breakdown of observed models.

(b) Self-predicting breakdown of observed models.

Figure 9: Breakdowns of models by the self-dominant and self-predicting conditions, needed to achieve truthfulness in OA and RPTS, respectively.

We start with the *self-dominant* condition, which is required for OA to have a truthful equilibrium:

$$P(s|s) > P(s'|s) \quad \forall s' \neq s. \tag{5}$$

A model is self-dominant if seeing a signal makes this signal the most likely signal for a peer. Figure 9a shows the breakdown. Most models do not satisfy this condition. An interesting observation is that it does not always hold even for binary models. This happens when one signal is much more likely than another: if an agent observes a very unlikely signal, she may still expect a peer to observe the more likely signal with probability more than 0.5.

Another condition is *self-predicting*, and is needed for the RPTS and related 1/prior mechanisms to have their intended properties:

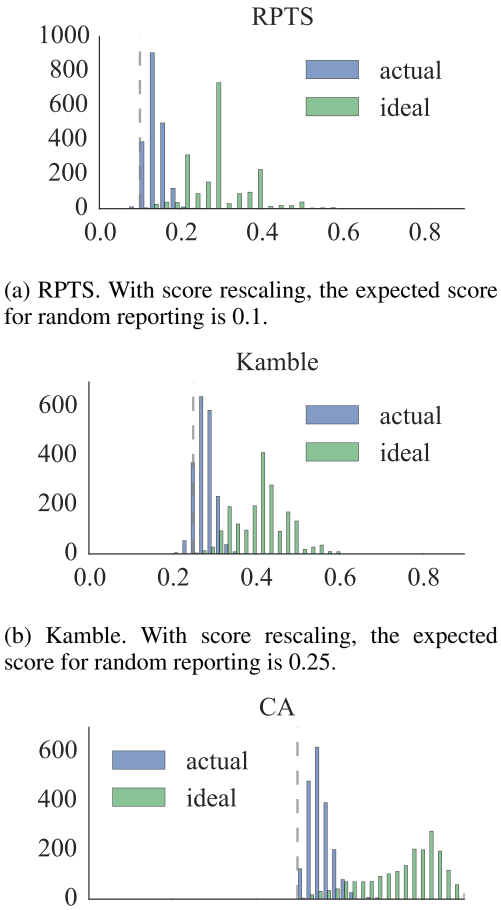
$$P(s|s) > P(s|s') \quad \forall s' \neq s. \tag{6}$$

In words, an agent’s peer is more likely to see a particular signal if the agent also sees that signal. Figure 9b shows the breakdown. This condition is weaker than the previous two, and holds for the majority of size three models, though not for most larger ones. This means that RPTS is manipulable in peer assessments with many options, though experiments would be needed to see whether students find the manipulations in practice.

6 Analysis II: Expected Score

As a second analysis step, we examine the incentives for investing effort in doing a careful assessment of a peer’s submission. For example, if a student can expect 50 out of 100

[9] As opposed to an unordered classification setting like labeling images as cars, animals, or people.



(a) RPTS. With score rescaling, the expected score for random reporting is 0.1.

(b) Kamble. With score rescaling, the expected score for random reporting is 0.25.

(c) CA. The expected score for random reporting is 0.5.

Figure 10: Histograms for expected scores vs. ideal scores with perfect agreement on reports, per mechanism. For all mechanisms, the relatively low agreement between student reports makes expected scores for truthfulness only slightly better than for random reporting, when compared to the overal range of possible scores. "Ideal" scores, if peers agreed perfectly, are much higher, so there is a need to improve agreement.

points by reporting randomly, and only 55 by carefully reviewing their peer, she may decide that the effort that careful review would take is not worth it. We omit OA from this analysis, because as discussed above, it is not strong truthful, and if enough students are willing to misreport, constant reporting will actually increase their scores.

We look at this numerically in Figure 10: the "actual" histograms show the expected scores for truthful reporting for all criteria. For all mechanisms, less than half the criteria have expected scores that are more than 0.05 above random reporting. The benefit to being honest is fairly small because of the noise in peer assessments.

To understand whether the small benefit from truthfulness relative to the working range of the scores is inherent (i.e., due to the non-uniform marginal distribution on reports) or due to the relatively low agreement between peers, we also plot the "ideal" expected scores that students would get in each mechanism, with the same signal prior but perfect agreement on reports. These are much higher, suggesting that there is an opportunity to address this problem by training students to peer assess more consistently; e.g., through better assessment rubrics, encouraging effort through schemes such as peer prediction, and through non-incentive-based methods (e.g. adding "Please do a good job. Your peers depend on it!" to the instructions), and compensating for student bias, for example using the methods described in Piech et al. (2013). Another pragmatic workaround may be to clamp scores more severely in order to expand the working range of scores.[10]

Finally, some peer assessment exercises are simply not appropriate for peer prediction: if submissions are judged very subjectively (e.g. "do you like this art by your peer?"), it would be better to ask for peer feedback and reward participation rather than trying to reward accuracy.

7 Analysis III: Variability

Most of the theoretical analysis in the peer prediction literature focuses on expected value, and says that agents prefer one strategy to another if the former has higher expected payoff.[11] However, variability in scores is also likely to be important for several reasons. First, fairness is important, especially in education: two students who do work of equal quality should get the same score. A second reason is risk aversion: a student whose expected score is 5 points will be happier to always get 5 points rather than a 25% chance of 20 points, and might prefer a more certain strategy with lower expected score. Finally, students are likely to learn better with consistent feedback.

Risk aversion is concerned with overall score variability, while for fairness, variance in scores *ex ante*, before seeing task, is ok—different types of tasks may reasonably give different expected scores.[12] We are more concerned about variance in score given a signal—as a student, if I assess two very similar submissions, give each the same score, but get very different feedback, I may feel cheated.

Since the mechanisms that we study have different effective score ranges, variance is not a good metric for comparison. Instead, we use the *coefficient of variation*; i.e., the

[10] However, this should be done with caution because it would break the incentive guarantees. For example, with RPTS, if we increase the minimal allowed prior to 0.25, then when a student got a signal that was less likely than 0.25, she could want to misreport, giving a more common response instead.

[11] One partial exception is Shnayder, Frongillo, and Parkes (2016), which uses replicator dynamics to model population learning rather than assuming equilibrium play based on expected rewards. The replicator dynamics evolve based on expected scores in a continuous population, so the core focus on expected score is still present.

[12] For example, in RPTS, unlikely reports have a higher expected score, so a student who gets a rare bad submission would expect more points than a students who gets a good submission.

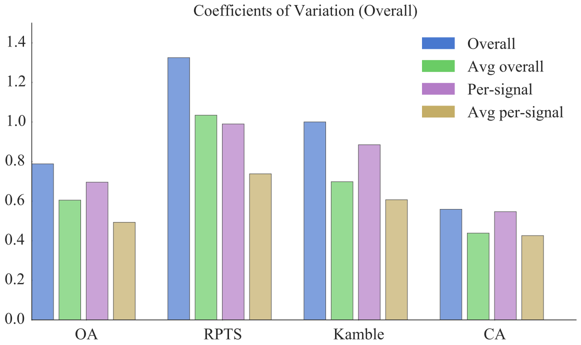


Figure 11: Overall coefficient of variation and expected per-report coefficient of variation of scores of different mechanisms, with and without averaging scores across all peers.

standard deviation divided by the mean. This is a standard way of comparing distributions with different scales.

Figure 11 shows the coefficient of variation for each mechanism, both overall and conditioned by signal. The signal-conditional value is the average of the individual coefficients of variation for each signal, weighted by the number of reports of that signal. In other words, it is the *a priori* expected coefficient of variation, before receiving a signal. Averaging scores for all peers always reduces the coefficient of variation, and the CA mechanism has the lowest overall variation.[13]

Figure 12 shows the details behind the averaged, per-signal bars in Figure 11. OA, RPTS, and Kamble all show similar patterns, because they are all based on output agreement and only adjust the relative payoff for each signal. The coefficient of variation for each k and signal is roughly inverse to the frequency of that report (recall Figure 6)—unlikely reports match less frequently, so get more varied scores. The coefficients of variation for CA are much more uniform across signals, because it rewards agreement between correlated signals as well as exact agreement.

Overall, CA appears to be better than the others candidate mechanisms in terms of variance. Unlike CA, RPTS, Kamble, and OA all rely on exact agreement, and so effectively work based on the difference between the diagonals of the joint and product-of-marginals distributions. As the number of signals goes up, there is less and less probability of two signals being exactly equal, making these mechanisms more fragile, relying on rare rewards for their guarantees, and increasing variance.

8 Analysis IV: Risk of Collusion

We now look at another potential problem with peer prediction. As Gao, Wright, and Leyton-Brown (2016) point out, students can potentially correlate on a *low-effort signal*, based on unintended and easy-to-observe properties of a submission such as length, id number, title, and so forth, thus matching without exerting effort. The suggested solution in Gao, Wright, and Leyton-Brown (2016) is to give up on peer prediction entirely, and use trusted TAs to spot check student evaluations. While that is certainly effective when TAs are available, we are working in a model without TAs, and look instead at the limits of what is possible under this kind of collusion. In particular, assuming that peer assignment is done randomly, we examine what fraction of the students needs to collude to benefit. It is likely in a large class that agreeing on such a correlation scheme would only be done by a fraction of the students.[14]

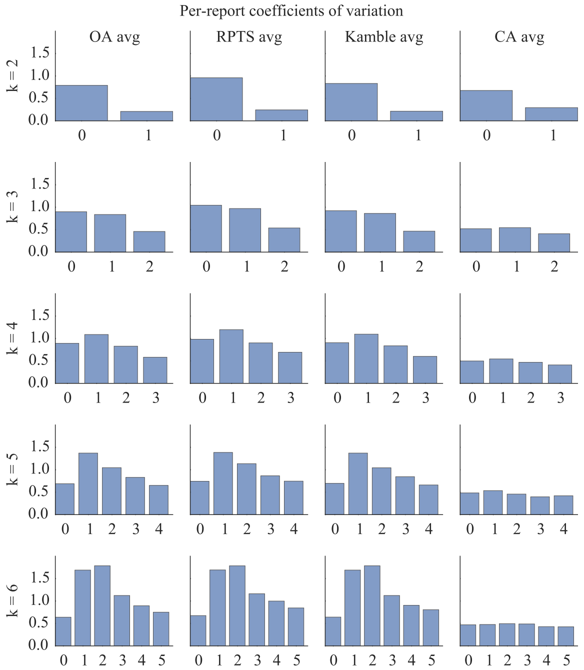


Figure 12: Per-report coefficients of variation for all mechanisms, averaging scores for all peers.

We focus on the CA mechanism here as the most promising candidate given the reward variance results above,[15] and assume a uniform distribution of low-effort signals, as in Gao, Wright, and Leyton-Brown (2016). Figure 13 shows the expected CA scores for a particular evaluation criterion chosen as an example, as the fraction of the population that is truthful varies, with the rest assumed to collude on a perfectly correlated low-effort signal. As expected given the ideal vs. actual score histograms in Figure 10c above, scores for the perfectly correlated low-effort signal are much higher than truthful scores. A large fraction of the population must

[13] RPTS has a bigger drop in variation from overall to per-signal, which makes sense because it uses different score ranges for different signals, so the overall distribution has high variance.

[14] We also note that there are reasons to expect collusion to be difficult in practice: students typically submit written feedback, not just score, and so still have to look at submissions. Students can complain if they get unfair evaluations, and students who are obviously cheating can be punished. Similarly, even a low percentage of spot checking can discourage cheating if the punishment is substantial.

[15] The results for Kamble and RPTS are similar.

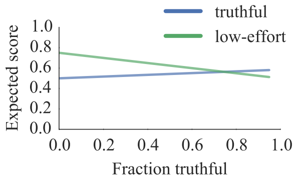


Figure 13: Example truthful vs low-effort scores using the CA mechanism, for a single criterion.

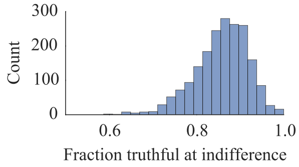


Figure 14: Fraction truthful at indifference across evaluation criteria.

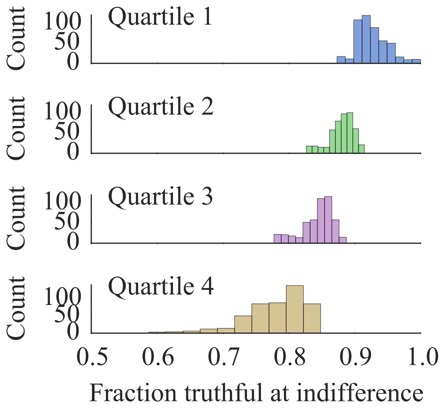


Figure 15: Histograms of the fraction truthful at indifference across evaluation criteria, with all criteria split into quartiles, sorted by amount of correlation. As correlation increases, the necessary fraction truthful goes down.

be truthful to get better scores than a subpopulation with perfect correlation. The intersection of the lines is the indifference point.

Figure 14 shows a histogram of the points of indifference for all the evaluation criteria. The values are quite high—80% or more of students have to be truthful for that to be the best strategy, given that the rest agree on a single perfect low-effort signal with uniform prior. The pattern is not very sensitive to how uniform the prior for the low-effort signal is, as long as it is not too extreme. It is quite sensitive to the score for truthful reporting, and to the assumption that all colluding students agree on a particular correlation method.[16]

The best solutions are to improve the likelihood of agreement for truthful reporting, which would make truthful scores go up and bring the indifference point lower, as well as spot checking and allowing complaints for low scores, as in *Mechanical TA* (Wright and Leyton-Brown 2015). To see the effects of improved agreement on the intended signal, we sort the assessment criteria by amount of correlation, as measured by total variation distance between the joint and product-of-marginals distributions (equivalently, the expected score of the CA mechanism). Figure 15 plots a separate histogram of the points of indifference for each quartile. The increased correlation in higher quartiles means a significantly lower point of indifference.

9 Conclusion

We examined patterns of reports in a MOOC peer assessment system, and simulated four peer prediction mechanisms applied to these reports. We found that agreement between peer reports is low overall, which raises some concerns about the use of peer prediction in this domain. On the other hand, we caution that the data comes from a system without incentives for accurate reporting. Incentive mechanisms are designed to boost effort and thus should improve agreement between reports. Ideally, the increased effort needed to be accurate will also improve student learning. Better agreement can also come through better student training and encouragement, and through bias-reduction techniques based on machine learning.

We argued that reward variance is an important consideration alongside expected score, for reasons of fairness, and find that the CA mechanism is better in this regard than mechanisms that only reward students based on exact agreement. An experimental follow-up question is whether the variability is low enough to be used in practice. A theoretical question is whether mechanisms with lower variability can be designed. We showed that collusion on unintended properties of submissions could be profitable with a small colluding sub-population, given the low base agreement rate, and suggest that improving agreement between peers and monitoring by the course team will help deter this behavior.

There are many directions for further research. In peer prediction, this includes exploring more mechanisms, perhaps using the information-theoretic framework from Kong and Schoenebeck (2016), which provides a general way to design mechanisms in a variety of settings. Another important direction is to find ways to handle user heterogeneity.

Our view is that these mechanisms are well enough understood that experiments and real deployments are both feasible and necessary to complement the theory. A concrete suggestion for a low-risk first implementation is to use peer prediction to give students feedback on how well they are assessing each other, without factoring the results into student grades. This should allow comparisons between mechanisms as well as an examination of the effects of ungraded feedback.

10 Acknowledgements

This research is supported in part by a grant from Google, the SEAS TomKat fund, and NSF grant CCF-1301976.

References

Busemeyer, J. R., and Townsend, J. T. 1993. Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment. *Psychological review* 100(3):432–459.

[16] It seems difficult to agree on such a method in practice in a large online course, without tipping off the course team by discussing it in some public forum.

Cho, K.; Schunn, C. D.; and Wilson, R. W. 2006. Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology* 98(4):891–901.

Dasgupta, A., and Ghosh, A. 2013. Crowdsourced Judgement Elicitation with Endogenous Proficiency. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*, 1–17.

Erev, I., and Barron, G. 2005. On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review* 112(4):912–931.

Erev, I., and Roth, A. E. 2014. Maximization, learning, and economic behavior. *Proceedings of the National Academy of Science (PNAS)* 111:10818–10825.

Falchikov, N. 1995. Peer feedback marking: Developing peer assessment. *Innovations in Education & Training International* 32(2):175–187.

Faltings, B.; Pu, P.; and Tran, B. D. 2014. Incentives to Counter Bias in Human Computation. In *The Second AAAI Conference on Human Computation & Crowdsourcing (HCOMP 2014)*, 59–66.

Gao, X. A.; Mao, A.; Chen, Y.; and Adams, R. P. 2014. Trick or Treat : Putting Peer Prediction to the Test. In *The Fifteenth ACM Conference on Economics and Computation (EC'14)*.

Gao, X. A.; Wright, R. J.; and Leyton-Brown, K. 2016. Incentivizing Evaluation via Limited Access to Ground Truth: Peer Prediction Makes Things Worse. *2nd Workshop on Algorithmic Game Theory and Data Science at EC 2016*.

Goldfinch, J., and Raeside, R. 1990. Development of a peer assessment technique for obtaining individual marks on a group project. *Assessment & Evaluation in Higher Education* 15(3):210–231.

Jurca, R., and Faltings, B. 2009. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research* 34(1):209–253.

Kamble, V.; Shah, N.; Marn, D.; Parekh, A.; and Ramachandran, K. 2015. Truth Serums for Massively Crowdsourced Evaluation Tasks. *arXiv:1507.07045*.

Kong, Y., and Schoenebeck, G. 2016. A Framework For Designing Information Elicitation Mechanism That Rewards Truth-telling. *arXiv:1605.01021*.

Kulkarni, C.; Bernstein, M.; and Klemmer, S. 2015. PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. *The Proceedings of the Second ACM Conference on Learning @ Scale*.

Kulkarni, C.; Wei, K. P.; Le, H.; Chia, D.; Papadopoulos, K.; Cheng, J.; Koller, D.; and Klemmer, S. R. 2013. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction* 20(6):1–31.

Miller, N.; Resnick, P.; and Zeckhauser, R. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51:1359–1373.

Orsmond, P., and Merry, S. 1996. The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education* 21.

Piech, C.; Huang, J.; Chen, Z.; Do, C.; Ng, A.; and Koller, D. 2013. Tuned Models of Peer Assessment in MOOCs. *Proceedings of the 6th International Conference on Educational Data Mining, Memphis, Tennessee*.

Prelec, D. 2004. A Bayesian Truth Serum For Subjective Data. *Science* 306(5695):462.

Radanovic, G., and Faltings, B. 2013. A Robust Bayesian Truth Serum for Non-Binary Signals. In *The Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13)*, 833–839.

Radanovic, G., and Faltings, B. 2015. Incentive Schemes for Participatory Sensing. In *AAMAS 2015*.

Radanovic, G.; Faltings, B.; and Jurca, R. 2016. Incentives for Effort in Crowdsourcing using the Peer Truth Serum. *ACM Transactions on Intelligent Systems and Technology* (January).

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

Shnayder, V.; Agarwal, A.; Frongillo, R.; and Parkes, D. C. 2016. Informed truthfulness in multi-task peer prediction. *The Seventeenth ACM Conference on Economics and Computation (EC'16)*.

Shnayder, V.; Frongillo, R.; and Parkes, D. C. 2016. Measuring performance of peer prediction mechanisms using replicator dynamics. *The 25th International Joint Conference on Artificial Intelligence (IJCAI-16)*.

von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, 319–326. New York, NY, USA: ACM.

Witkowski, J., and Parkes, D. C. 2012a. A Robust Bayesian Truth Serum for Small Populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012)*.

Witkowski, J., and Parkes, D. C. 2012b. Peer Prediction Without a Common Prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC2012)*.

Wright, J. R., and Leyton-Brown, K. 2015. Mechanical TA : Partially Automated High-Stakes Peer Grading. In *SIGSCE'15*.