Calibrated Self-Assessment

Igor Labutov
Cornell University
iil4@cornell.edu

Christoph Studer
Cornell University
studer@cornell.edu

ABSTRACT

Peer-grading is widely believed to be an inexpensive and scalable way to assess students in large classroom settings. In this paper, we propose *calibrated self-grading* as a more efficient alternative to peer grading. For self-grading, students assign themselves a grade that they think they deserve via an incentive-compatible mechanism that elicits maximally truthful judgements of performance. We show that the students' self-evaluation scores obtained via this mechanism can be used to perform classic item response theory (IRT) analysis. In order to obtain unbiased estimates of the IRT parameters, we show that the self-assigned grades can be calibrated with a minimum amount of input from instructors or domain experts. We demonstrate the effectiveness of the proposed calibrated self-grading approach via simulations and experiments on Amazon's Mechanical Turk.

Keywords

Assessment, self-grading, item response theory (IRT).

1. INTRODUCTION

A significant bottleneck in scaling traditional classrooms to hundreds or thousands of students is the challenge of enabling efficient mechanisms of assessment. Peer-grading, hailed as a solution to this "scaling problem," has received significant attention, both from the education [12, 5] and machine learning [10, 11] communities. Broadly speaking, peer-grading can be thought of as a relaxation of the traditional teacher/student roles in the classroom: An expert instructor is replaced by several "noisy" students having the task of estimating performance of other students. Virtually all of the existing statistical models for peer-grading aim to estimate the student's true performance from such noisy measurements, under some metric of optimality.

Self-grading constitutes a special case of peer-grading: The student is their own only "peer" and is solely responsible for assigning a score based on the judgement of their own work. Depending on the student's honesty in self-evaluation, self-grading is appealing for at least two reasons: (i) Students can provide a richer signal towards their internal state of knowledge by explicitly revealing confidence in their answers—a signal that can be exploited during assessment; (ii) because every student is their own grader, potentially no additional peer-grading efforts are required to perform assessment. Self-grading, however, introduces two unique challenges not faced in traditional peer-grading: (i) Designing mechanisms for eliciting honest judgement of performance and (ii) accounting for individual biases in self-evaluation. The first challenge in self-grading fundamentally requires an explicit mechanisms for eliciting truthful judgements.[1] The second challenge is addressed in peer-grading by appealing to statistics and assuming that the population of graders is—at least on average—unbiased.

In this work, we propose *calibrated self-assessment* to address both of the above challenges. Our approach combines self-assessment with a small number of instructor-graded items, which provides a simple, incentive-compatible mechanism of eliciting self-assigned scores, and yields assessments of comparable or superior quality to a setting with significantly more instructor-graded items and no self-scoring. As a consequence, calibrated self-assessment enables a significant reduction in effort of instructors, domain experts, or peers.

2. RELATED WORK

We focus our review on two research directions that our work aims to bring together: (i) self-assessment as a method for summative assessment and (ii) decision-theoretic mechanism design for judgement elicitation.

Self-grading and Peer-grading in education: Self-assessment is often seen by teachers as a valuable tool in classrooms [17], who cite self-assessment as a viable way to reduce the instructor's effort, elicit additional information from students (e.g., their effort and confidence), and provide an additional learning opportunity in the process. More recently, in addition to peer-grading, self-grading was deployed in massive open online courses (MOOCs) [5]. Self-grading as a tool for summative assessment, however, is controversial, with its validity questioned on the basis of students' internal biases. In fact, studies indicate that bias is often a function of one's ability [17, 16]. Studies that compare peer-grading and self-grading differ in their findings, with self-grading and

[1] This is also a potential problem in peer-grading when conflicts of interest are present.

peer-grading performance excelling in different conditions (classrooms, age-groups, etc.), but both are heavily influenced by the underlying assessor biases (see [16] for a survey of the studies). A study carried out in four middle-school science classrooms found that peer-grading and self-grading have a high correlation with instructor grades, with grading bias patterns that are consistent with other studies [12]. In addition, they found that the process of self-grading resulted in learning gains, whereas peer-grading did not. A recent study carried out at the university level, however, found that both peer-grading and self-grading results in learning gains as a side-effect of grading [8].

The existing literature on self-grading points to the significant effect of bias in self-scoring, with most studies concluding that students of lower ability tend to inflate their grades more. As a consequence, we argue for the importance of an incentive-compatible mechanism that is designed to elicit maximally truthful judgements, and a *calibrated* model that is able to explicitly de-bias the individuals by incorporating a subset of instructor-graded items.

Judgement elicitation: The literature on truthful judgement elicitation through scoring functions dates back to the fifties, when the so-called “quadratic scoring rule” was proposed for the task of weather forecasting [2]. Since then, a number of generalizations of the quadratic scoring rule and other incentive-compatible scoring rules have been proposed and analyzed [3, 14, 7, 13] and found application in forecasting weather, sports, and finance. Analysis of the behavior of non-risk neutral agents in scoring-rule-based mechanisms has received only limited attention [9], with lottery-based payoffs being the most well-known solution for encouraging risk-neutral behavior. Lottery-based payoffs had received mixed results in experimental evaluations [4, 15], and in the context of education a reward system based on a lottery is not a reasonable solution. In this work, we rely on heavily limited instructor input in order to correct for individual biases, which includes under- and over-confidence, as well as non-risk-neutral behavior.

To the best of our knowledge, the only work that applies a scoring rule mechanism in the context of education that we are aware of is [1]. The focus of this work is in analyzing the effect of different scoring functions on the self-assessment behavior of students. Our primary contribution in this work is in developing a principled statistical model for calibrated summative assessment that integrates self-scoring and instructor-scoring within the classic IRT framework.

3. MODEL

Self-grading without a proper incentive mechanism may lead to dishonest behavior. In the setting of self-grading, a “mechanism” is a *scoring rule* that specifies the rules by which the points are assigned to the student as a function of their own judgement and the outcome (i.e., whether their answer was correct). A mechanism is called *incentive compatible* when the student's optimal strategy with respect to his or her own utility function results in a truthful elicitation of information, e.g., truthful judgement of their own work.

We consider the following scoring function:

$$p_{ij} = \begin{cases} \theta_{ij} & \text{if correct} \\ -\frac{1}{2}\theta_{ij}^2 & \text{if wrong,} \end{cases}$$

where $\theta_{ij} \in [0, A]$ is a score provided by student i in answering question j, where A is some fixed upper bound. If the student provides a correct answer, they get the θ_{ij} points that they proposed; if they provide an incorrect answer, they lose exactly half of that value squared. This scoring function is known as a *quadratic scoring rule* and was first proposed in [2].

For this scoring function, the expected payoff is

$$\mathbb{E}[p_{ij}] = \theta_{ij}\hat{\pi}_{ij} - \frac{1}{2}\theta_{ij}^2(1 - \hat{\pi}_{ij}), \tag{1}$$

where $\hat{\pi}_{ij}$ is the i^{th} student's estimate of the probability that they will get question j correct. This expression is maximized when

$$\theta_{ij} = \frac{\hat{\pi}_{ij}}{1 - \hat{\pi}_{ij}}. \tag{2}$$

Equation 2 is exactly the student's own belief about the odds of them answering the question correctly. Consider that the student estimates their chances of answering any question correctly, by simultaneously estimating their own ability and the difficulty of the question. Let us now define that probability to be the standard IRT Rasch likelihood, but defined with respect to the student's own estimate of their ability, $\hat{s}_i$ and their estimate of the question's difficulty $\hat{q}_j$:

$$\hat{\pi}_{ij} = \frac{1}{1 + \exp(-(\hat{s}_i - \hat{q}_j))}.$$

Given the student's estimate of their own ability $\hat{s}_i$ and of the difficulty of the question $\hat{q}_j$, we can now derive their optimal proposed score (assuming they act rationally and are risk-neutral) for that problem θ_{ij} (or rather its logarithm):

$$\log(\theta_{ij}) = \hat{s}_i - \hat{q}_j,$$

which follows from the fact that log-odds of a logistic model is a linear function of its parameters. We will assume that the student is risk-neutral and is unbiased in his or her estimates of own ability and question difficulty, but we will relax both assumptions later. On any given question, however, the student's estimate of their ability to answer that particular question may deviate from their true ability. Assuming that the student's own estimates are normally distributed around their true values, we get:

$$\hat{s}_i - \hat{q}_j \sim \mathcal{N}(s_i - q_j, \sigma^2),$$

where s_i and q_j are the true student ability and question difficulty parameters respectively. As a consequence, it follows that $\log(\theta_{ij})$ is normal distributed and θ_{ij} is log-normal distributed. Consider a dataset D consisting of the self-assigned scores $\log(\theta_{ij})$ submitted by each student for each question that the student answered. We can write the conditional likelihood of the entire dataset as follows:

$$P(\boldsymbol{\theta} \mid \mathbf{s}, \mathbf{q}) = \prod_{(i,j)\in D} \mathcal{N}(\log(\theta_{ij}) \mid \mu = s_i - q_j, \sigma^2).$$

Here, $\mathbf{s}$ and $\mathbf{q}$ are the vectors comprising the student ability and question difficulty parameters, respectively, and $\boldsymbol{\theta}$ is the vector of student-submitted scores. Maximizing the likelihood of all observations gives a straightforward least-squares solution for the parameters s_i and q_j, given all the user-provided scores θ_{ij}. Note that σ^2 is assumed to be a constant variance in students' estimates of their own ability. In practice this variance is likely user-specific and corresponds to the students' ability in self-assessment. We will address the issues of bias and variance in self-assessment in Section 3.2.

3.1 Parameter estimation

It is interesting to note that we can solve for the IRT parameters (student abilities and question difficulties) using the above formulation with *no* outcome information, i.e., without knowing which students answered which questions correctly. In fact, the above approach does not even require that the students who are self-grading know what the correct answer is; students' confidence in their answers elicited through the quadratic scoring rule is all that is needed to learn the parameters of the model. Of course, this observations relies on two fundamental assumptions: (i) students are risk-neutral and (i) students are unbiased in estimating their chance of answering a question correctly. In Section 3.2, we will account for the individual biases and non-risk-neutral behavior by explicitly introducing a bias parameter into the model and estimating it from an additional set of instructor-graded responses. However, in order to gain a better understanding of the model, it is insightful to first analyze the solution to the problem where both of these assumptions hold.

The solution for the model parameters can be obtained in closed-form using a standard pseudo-inverse solution to a least-squares problem. Alternatively, the solution can be obtained iteratively, without requiring to explicitly invert any (potentially large) matrices. In particular, one can repeatedly evaluate the following two steps:

$$\begin{aligned} s_i &= \sum_{j \in Q_i} \frac{q_j}{\lambda + n_q^i} + \sum_{j \in Q_i} \frac{\log(\theta_{ij})}{\lambda + n_q^i} \\ q_j &= \sum_{s \in S_j} \frac{s_i}{\lambda + n_s^j} - \sum_{i \in S_j} \frac{\log(\theta_{ij})}{\lambda + n_s^j}. \end{aligned}$$

Here, s_i is the ability of student i and q_j is the difficulty of question j. To guarantee a unique solution, we introduce a non-negative regularization parameter λ, which we will discuss in more detail in the next paragraph. The constants n_q^i and n_s^j are the number of questions that student i answered and the number of students that answered question j respectively. Note that the above iterative solution has an intuitive interpretation: The ability of the student is the sum of the average of the (log-transformed) self-assigned scores to a set of questions that the student answered and the average difficulty of those questions. In turn, the difficulty of a question is the negative of the average (log-transformed) score that students assigned to themselves for that question plus the average ability of the students who answered that question. Intuitively, if students with high ability self-assess themselves to have done poorly on a specific question, that question will have a large difficulty parameter.

In the case where there is no missing data, i.e., each student answers each question, the solution for student ability parameters simplifies to:

$$\mathbf{s} = \begin{bmatrix} \frac{\sum_{i \in S} \log \theta_{i1}}{\lambda + N_s} \\ \vdots \\ \frac{\sum_{i \in S} \log \theta_{iN_q}}{\lambda + N_s} \end{bmatrix} + \mathcal{O}(1/\lambda)\mathbf{1},$$

where $\mathcal{O}(1/\lambda)$ is a function that grows proportional to $1/\lambda$. In other words, the student's ability is simply the average of the (log-transformed) scores that the student assigned to themselves plus a constant that is identical for each student. This solution also illustrates the role of the regularization parameter λ. Because the solution for $\mathbf{s}$ and $\mathbf{q}$ is location-invariant, without an explicit prior, the likelihood is maximized by scaling all parameters to infinity. This is equivalent to setting λ to 0, in which case the above solution will tend to infinity, as expected. Note, however, that the relative ranking of the student abilities in this solution will be consistent, regardless of λ. As obtaining the ranking of the students is our primary focus, we can thus set λ to zero in the above solution, and simply consider the average self-assigned (log-transformed) score as the the ability parameter of the student. The same argument applies to question difficulty parameters.

3.2 Calibrating the model

There are two issues in relying on students' self-given score for ranking students via the IRT model: (i) Students may be prone to over- or under-estimating their ability and (ii) because there is uncertainty involved in both answering and grading, some students may be more or less inclined to "gamble" with their self-assigned score (i.e., some students are more or less risk-averse/risk-loving). We subsume both effects (as it is impossible to tell them apart) into a general student "bias" in self-grading, and model it explicitly as

$$\log(\theta_{ij}) = \hat{s}_i - \hat{q}_j + b_i,$$

where $b_i \in (-\infty, \infty)$ is a student-specific bias. We assume that this student bias is drawn from a normal distribution $b_i \sim \mathcal{N}(0, \sigma_b^2)$, where the above distribution stipulates that the average of the student population is unbiased. It is impossible to estimate b_i using self-grading alone, as without actual observations of correctness of students' responses, the model will conflate s_i and b_i into a single parameter. Imagine that we do grade a student's responses on a small subset of the answered questions (which they also self-grade). Let the set of instructor-graded questions be $Q_g \subseteq Q$, where Q is the set of all questions. As the observations of instructor- and self-assigned grades are all conditionally independent given the student and question parameters, the overall likelihood of both self- and instructor-given scores is a product of these likelihoods. We can then express the log-likelihood of the entire dataset as a sum of the self-graded response log-likelihoods and instructor-graded response log-likelihoods:

$$\begin{aligned} \log P(\boldsymbol{\theta}, \boldsymbol{y} \mid \mathbf{s}, \mathbf{q}, \mathbf{b}) = \sum_{s_i \in S} \Bigg(& \underbrace{\sum_{q_j \in Q} (\log \theta_{ij} - (s_i + b_i - q_j))^2}_{\text{self-graded responses}} \\ & + \underbrace{\sum_{q'_j \in Q_g} \log(1 + \exp(-y_{ij}(s_i - q'_j)))}_{\text{instructor-graded responses}} \Bigg). \end{aligned}$$

Here, $y_{ij} \in \{-1, 1\}$ is the instructor-grade for question j answered by student i and $\boldsymbol{y}$ is the response vector for all students ($y_{ij} = +1$ corresponds to a correct response and $y_{ij} = -1$ otherwise). Observe that the "bias" parameter only appears in the self-graded part of the likelihood. This allows us to calibrate the model via instructor-graded questions as a "training set" to separate the effects of the bias and true ability. Note that, unlike in the previous case that relied entirely on students' self-scores, like with the traditional Rasch IRT model, we are unaware of a closed form solution for this formulation. In all of our experiments, we use the L-BFGS algorithm [18] for learning model parameters.

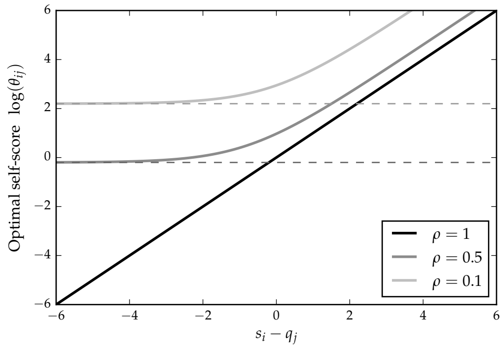


Figure 1: The optimal strategy for providing a self-assessment score $\log \theta_{ij}$ for a student with ability s_i on a question of difficulty q_j, assuming the student's knowledge that a random fraction ρ of the questions will be graded. The optimal strategy is approximately piece-wise linear as a function of the student's relative ability $s_i - q_j$. In the regime of low relative ability, the student's optimal strategy is to report a fixed score that is a function of ρ, regardless of his or her relative ability.

3.3 Consequences of students' awareness of the mechanism

The assumption that the learner is optimizing a utility function based on the expected test score:

$$\mathbb{E}[p_{ij}] = \theta_{ij}\hat{\pi}_{ij} - \frac{1}{2}\theta_{ij}^2(1 - \hat{\pi}_{ij}) \tag{3}$$

fundamentally assumes that the student believes that each question will be graded, as otherwise there would be no possibility of getting a question wrong and losing points. In practice, our goal for self-grading may be motivated by the effort to reduce the instructor's involvement in grading, and, in general, as a way to scale assessment to potentially very large classrooms, such as massive open online courses (MOOCs). Having each submission be graded by an instructor (or your peers) defeats the purpose of self-grading. If, however, the student is aware of the fact that not every question is graded, we can expect that their utility function, and thus their optimal strategy, will be affected by this knowledge. If the test is administered once, of course, the students could be deceived into believing that every question is graded. In a real course, however, a more realistic assumption is that the students possess the knowledge that not all of the questions are graded and if the assignments are returned, we can expect that the students' estimates of the fraction of graded questions will improve over time. If, however, the student believes that a random subset of their submissions is graded by someone else, but if the student does not know which subset is graded, then we should still expect the student's optimal behavior to be maximizing a utility function similar to the one above. The utility function will not be the same, as we now have to account for the student's belief about how many problems are graded by someone else. Let us assume that the student has a prior belief that each problem has a probability ρ of being graded. Then, the expected score the student i receives on question j is given by

$$\mathbb{E}_{gr}\left[\mathbb{E}[p_{ij} \mid \text{graded}]\right] = \rho(\theta_{ij}\hat{\pi}_{ij} - \frac{1}{2}\theta_{ij}^2(1 - \hat{\pi}_{ij})) + (1 - \rho)\theta_{ij},$$

where we take an additional expectation with respect to the student's belief that the problem is graded. Note that when a problem is *not* graded, the expected score that the student receives is just θ_{ij}, i.e., their self-assigned score, regardless of whether the student answers correctly. This is because when a problem is not graded, there is no possibility of losing points. We can show that the student's optimal self-assigned score $\log(\theta_{ij})$ has the following approximate relationship to their ability and question difficulty (the approximation is a piece-wise linear approximation to the true strategy that is asymptotically accurate):

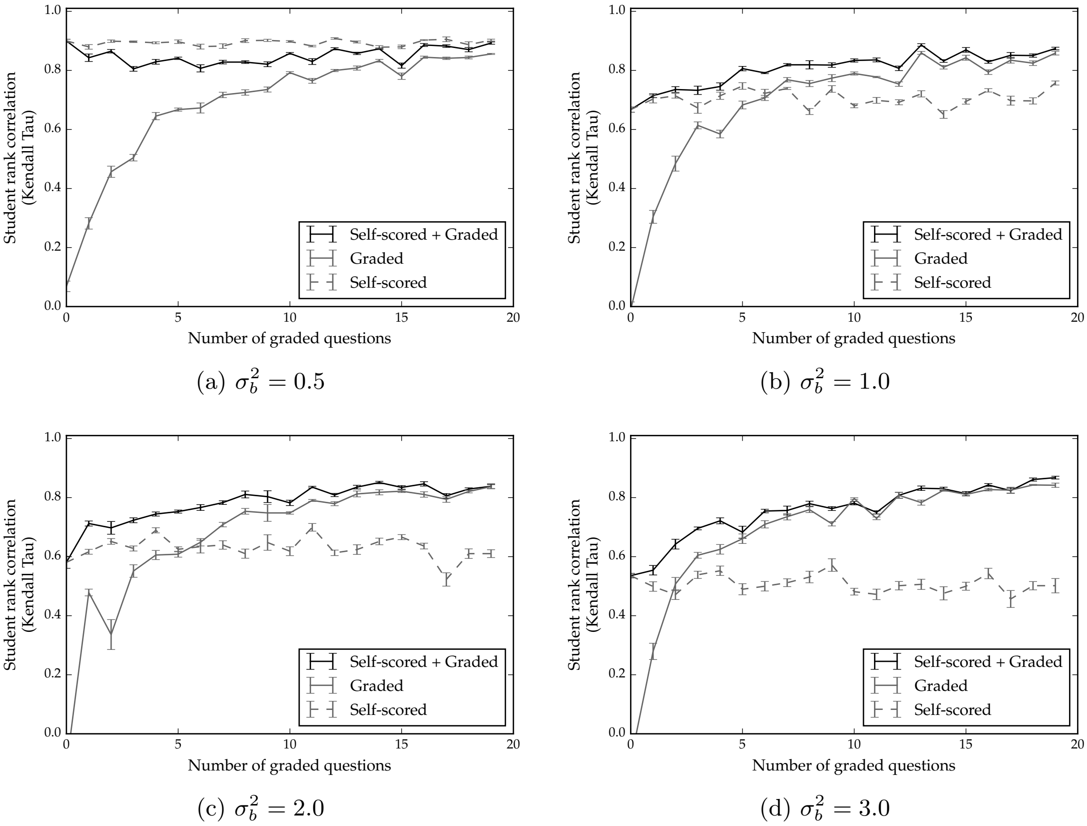
$$\log(\theta_{ij}) = \max\left\{\log\left(\frac{1}{\rho} - 1\right), (s_i - q_j) - \log\rho\right\}.$$

The optimal strategies for different values of ρ are illustrated in Figure 1. The student's knowledge of the mechanism is reflected by the appearance of a lower-bound on the self-assigned score in a region where the student is likely to do poorly (low values of $s_i - q_j$). This is expected: If the student is aware that the chance of a particular question to be graded is low enough, it would make sense to take advantage of those odds and "bet" a small, but a non-zero amount, even if the student does not know the correct answer. From a practical perspective of implementing a system that solicits self-assessment scores, it would not make sense to provide the user with the ability to provide a self-assessment score lower than their optimum. From the model inference perspective, this introduces a complication: Observations that correspond to the lowest possible self-score do not correspond to any specific $s_i - q_j$, but rather an entire range. This problem is known generally as *censored regression.* and can be solved using the same approach as for the original problem, but with the modified likelihood function that accounts for this "kink." Note that a similar restriction on the likelihood (but as an upper-bound) is introduced when the maximum attainable score for a problem is incorporated into the scoring function.

4. EXPERIMENTS

4.1 Simulations

It is insightful to study the effect of bias in the population of students on the quality of the learned parameters in the IRT model: student ability parameters and question difficulty parameters. We perform a simple simulation of a classroom with 50 questions and 30 students (question difficulties and student abilities are sampled from a zero-mean normal distribution with a standard deviation of 3), where each student answers each question (a total of 1,500 responses). In this simulation, each student submits their self-grade $\log(\theta_{ij})$ for



(a) $\sigma_b^2 = 0.5$

(b) $\sigma_b^2 = 1.0$

(c) $\sigma_b^2 = 2.0$

(d) $\sigma_b^2 = 3.0$

Figure 2: Simulation results. Rank correlation across students obtained using three models for different variance of self-grading bias (σ_2): (i) *black*: a model that uses student self-scores and the correctness of their response to a subset of graded questions (number of graded questions on x-axis), (ii) *solid gray*: a model that uses correctness of their response to a subset of graded questions only (number of graded questions on x-axis) and (iii) *dashed gray*: a model that uses only the students' self-score.

each question by optimizing their utility according to the utility function in 3. We repeat the simulation for four different populations of students, each with a different variance σ_b^2 of the bias parameter. To evaluate the quality of the inferred student parameters, we compute the rank correlation (Kendall Tau) between the true ordering of the students (by their true parameters) and the ordering obtained by sorting the students based on the inferred parameters. The Kendall Tau metric is defined as follows:

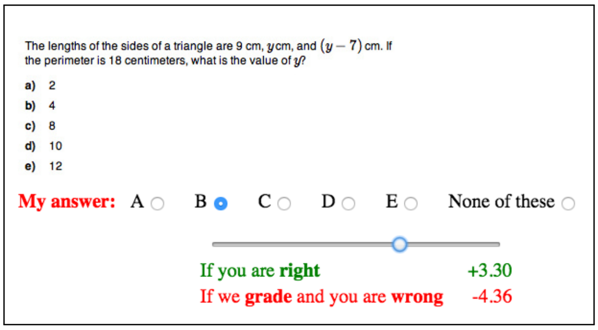
$$KendallTau(\mathbf{s}, \hat{\mathbf{s}}) = \frac{N_{\text{pairs}}^{\text{correct}} - N_{\text{pairs}}^{\text{wrong}}}{N_{\text{pairs}}}$$

where $\mathbf{s}$ and $\hat{\mathbf{s}}$ are the true and inferred student ability parameters, respectively, and $N_{\text{pairs}}^{\text{correct}}$ and $N_{\text{pairs}}^{\text{wrong}}$ is the number of student pairs that are ordered correctly in the inferred ranking (with respect to the true ranking) and the number of pairs that are ordered incorrectly, respectively. Kendall Tau is equal to $+1$ when the rankings are consistent and to -1 when the rankings are inverted. The corresponding results are shown in Figure 2.

Three models were evaluated:

- **Self-grading only**: Only students' self-submitted scores $\log(\theta_{ij})$ are used in fitting the Rasch model parameters. All students submit their self-scores for all questions. The correctness of students' responses is not used in fitting the Rasch parameters.
- **Instructor-grading only**: Only the correctness of the responses is used for fitting the Rasch model parameters; this is a classic Rasch model. We vary the number of questions used in fitting the model parameters (x-axis in Figure 2).
- **Self-grading + instructor-grading**: A combination of self-scores submitted by all students for all questions and the correctness of a subset of submitted questions is used for fitting the Rasch model parameters (number of questions used is the x-axis in Figure 2).

In the case where the students in the class are relatively unbiased (low σ_b^2) (top left in Figure 2), self-scoring achieves a better rank-correlation than the traditional IRT Rasch model, even when many questions are instructor-scored. Interestingly, in the regime of low bias, including actual instructor-graded responses actually negatively affects the correlation (this is due to over-fitting caused by a small number of instructor grades—introducing additional bias variables requires a sufficient number of observations to infer them reliably; this performance drop eventually disappears when a sufficient



The lengths of the sides of a triangle are 9 cm, y cm, and $(y - 7)$ cm. If the perimeter is 18 centimeters, what is the value of y?

a) 2
b) 4
c) 8
d) 10
e) 12

My answer: A ○ B ● C ○ D ○ E ○ None of these ○

If you are **right** +3.30
If we **grade** and you are **wrong** -4.36

Figure 3: Screenshot of one question from the Mechanical Turk task. A subject answers a math question and provides a self-assessment score by adjusting a slider. The student sees the number of points that they will gain if they answer the question correctly (green) and the number of points they will lose if they answer the question incorrectly (red).

number of questions is included). As the bias of the population increases, the performance of the self-scoring model decreases but still exceeds the performance of the instructor-only Rasch IRT, especially in situations where only a few questions are scored.

4.2 User study

To evaluate the efficacy of the proposed self-grading approach, we conducted a user-study on Amazon's Mechanical Turk. We solicited 206 subjects to participate in a task titled "Do a short math quiz and earn bonus!". The subjects were asked to answer 30 math questions of varying difficulty levels ranging from basic arithmetic to pre-calculus. The questions from the dataset introduced by [6] were used in our experiment. All questions were multiple choice and included a "none of the above" option, included in order to minimize the probability of getting a right answer through a process of elimination. Although in practice, multiple-choice questions mostly defeat the purpose of self-grading, we use multiple choice questions for the ease of evaluation and the lack of subjectivity that would be otherwise present in free-response questions. Figure 3 illustrates a single question from the task. The subjects were asked to mark what they believed to be the correct answer, and then to assign themselves the number of points that they would receive if they answered the question correctly. The input was provided through a slider. Moving the slider automatically displayed the number of points that the subject would gain if they answered the question correctly (green), and the number of points they would lose if they answered the question incorrectly (red). The points were then converted to currency (1 point = \$0.01), and paid through a "bonus" mechanism in Mechanical Turk. We chose to use real currency as a reward to to ensure that the subjects had a stake in their performance, and thus there is incentive to think carefully about their self-assigned scores.

We follow the same evaluation scheme that we described in the previous section. Recall, that we are interested in the quality of the assessment derived from the students' self-evaluation. In the simulation study, a "gold-standard" assessment was available and allowed us to use rank correlation between the "gold-standard" ranking and the inferred ranking as an evaluation metric. In this user-study, we consider the ranking inferred by the IRT model that relies on the complete dataset, as a proxy for the "gold-standard" ranking. We then repeat the evaluation scheme described in the previous section: (i) vary the number of instructor-graded questions from 0 to all questions (30) and combine that with the self-assigned scores for every question, (ii) infer the ranking using the proposed model, and (iii) compare it to the ranking that is derived from "gold-standard" proxy.

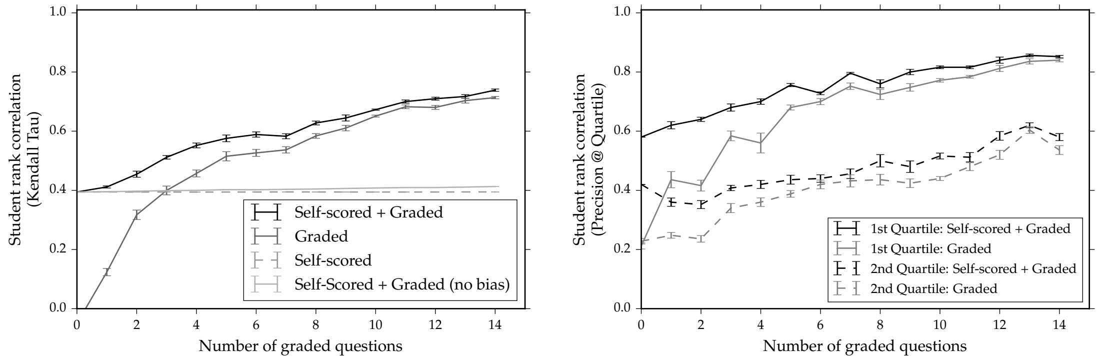
We find that the results are comparable to those obtained in the simulation (Figure 4(a)). Self-scoring is already able to obtain a reasonable correlation with the "gold-standard" ranking even without any instructor-graded question. Incorporating instructor-grades for additional questions improves the performance. Rank correlation metrics, such as Kendall Tau, while convenient for summarizing the results with a single quantity, often fail to distinguish regimes where the model might perform differently. It is instructive to consider the performance of rank-correlation in the different segments of the ranking. Figure 4(b) decomposes the results by quartiles. We employ a more intuitive metric, *Precision@Quartile*, defined as follows:

$$\mathit{Precision@}Q_i = \frac{|\hat{S}_{Q_i} \cap S_{Q_i}|}{|\hat{S}_{Q_i}|}$$

where S_{Q_i} is the set of students in the ith quartile of the "gold-standard" ranking, and $\hat{S}_{Q_i}$ is the set of students in the ith quartile of the inferred ranking. This metric captures the ability of the model to perform within a particular segment of the ranking. For example, looking at Precision at the first quartile, measures the ability of the model to predict top students. From Figure 4(b) we can conclude that the model is significantly better at distinguishing the top-ranked students (first quartile) as compared to the lower-ranked students (second quartile). By using the self-scoring signal without any instructor-graded questions, we are able to recover nearly 60% of the top quarter of all students. The performance in the second quartile is significantly lower, but follows the same trend: incorporating the students' self-reported scores in the regime of zero to several questions significantly improves performance over the baseline of instructor-graded questions alone. This observation leads to the conclusion that, at least in this study, better students were better at estimating their ability. We look into the effect of self-estimation performance in more detail in the next section.

4.3 Self-assessment and bias

The performance of the model that relies on self-assessment depends fundamentally on the model's estimates of the students' biases as well as the ability of the students to self-assess reliably (self-assessment variance). In our model, we infer only the individuals' biases and assume constant variance in the self-assessment likelihood (these could in principle be estimated as well. Figure 6 illustrates the individual inferred biases for each student (averaged across multiple folds), sorted in an increasing order. The resulting distribution illustrates the skew in the bias distribution towards "under-confidence," i.e., most students tend to under-estimate their ability (act conservatively). The importance of estimating bias is underlined in Figure 4(a), where we include an additional baseline **Self-Scored + Graded (no bias)** (light solid line). This baseline combines self-assessment and instructor-grades but does *not* incorporate the explicit student-bias parameter. As



(a)

(b)

Figure 4: User study results. Rank correlation across students obtained using three different models (i) **Self-scored**: a model that relies entirely on student-submitted self-assessments, (ii) **Graded**: a model that relies entirely on instructor-provided grades, as a function of the number of graded questions (x-axis), and (iii) **Self-scored + Graded**: a model that aggregates students' self-assessment scores on all questions and a variable number of instructor-graded questions (x-axis). (a) Computes rank correlation across all students using Kendall Tau, and (b) decomposes rank correlation across the first two quartiles using the *Precision@Quartile* metric. The model that combines self- and instructor-assigned scores is significantly better at predicting the top-performing students (first quartile). Combining instructor grades with self-assessment significantly improves both rank measures, especially when only a few questions are graded. Note that the total number of questions in the study was 30; we display the results up to 15, as the differences between both models is not substantial beyond that.

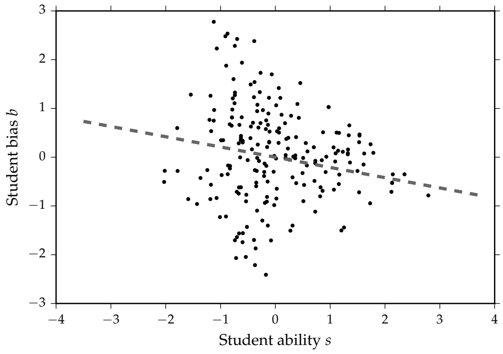


Figure 5: Bias vs. ability (centered). Both parameters were inferred using all of the available data. Each point in the scatter-plot corresponds to one student. A weak, but significant correlation between bias and ability exists.

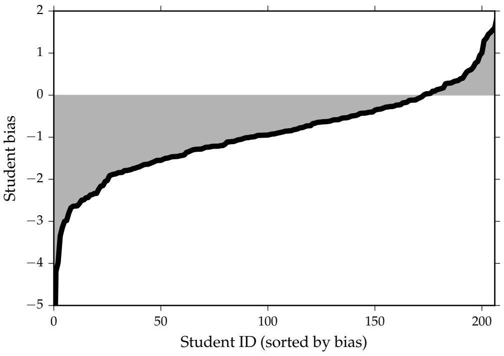


Figure 6: Inferred bias parameter of each student (sorted in an increasing order). The bias parameter was inferred using all of the available data.

evident from the graph, estimating bias is critical for combining self-grading and instructor-grading: without the bias parameter, the model is not able to leverage the benefits of both signals.

It is potentially insightful to investigate the relationship between self-assessment bias and ability. We consider the inferred bias parameter after incorporating instructor-grades for all questions, and compare it to the inferred ability parameter of each student. The result is illustrated in the scatter-plot in Figure 5. While the relationship between the two is not strong, there exists a negative correlation between ability and self-assessment bias (Pearson's correlation: 0.17, p-value = 0.013). Students that are more able tend to underestimate their ability, and students that are less able tend to inflate their ability. This finding is consistent with the literature in self-assessment [17, 16].

5. CONCLUSION AND FUTURE WORK

In this work, we have developed a novel approach for performing calibrated, summative self-assessment by combining (i) student's self-evaluations obtained via an incentive-compatible scoring mechanism and (ii) a minimal number of instructor-graded responses. We have shown that when the scoring rule is quadratic, the standard IRT Rasch model reduces to standard linear regression. We have demonstrated that the quality of the inferred assessment using self-scoring alone without additional instructor input is, on-average, comparable to the performance obtained using the standard IRT that requires significant instructor effort. Furthermore, by incorporating a minimum number of instructor-graded responses, we have shown that our approach substantially improves the estimates of the students' abilities and the questions' difficulties. Finally, we have addressed the longstanding issue of applying scoring rules in practice: dealing with the consequences of individuals' biases and non-risk-neutrality. We have proposed to explicitly model the combined effect of these two factors within the standard IRT framework, allowing the model to effectively de-bias these individual differences.

Our results open an interesting direction of inquiry: are there other scoring functions that are more efficient at estimating IRT parameters, and if so, can the scoring functions be adapted to individual students and questions, improving the efficiency of adaptive testing? In order to facilitate further research in this direction, we release all code and data used in this study.

6. ACKNOWLEDGMENTS

We would like to thank Mr. Lan, Ph.D., as well as A. E. Waters and R. G. Baraniuk for their help with the Mechanical Turk dataset [6]. The work of I. Labutov was supported in part by a grant from the John Templeton Foundation provided through the Metaknowledge Network at the University of Chicago. The work of C. Studer was supported in part by Xilinx Inc. and by the US NSF under grants ECCS-1408006 and CCF-1535897.

7. REFERENCES

[1] J. E. Bickel. Scoring rules and decision analysis education. *Decision Analysis*, 7(4):346–357, 2010.

[2] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[3] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[4] G. W. Harrison, J. Martínez-Correa, and J. T. Swarthout. Inducing risk neutral preferences with binary lotteries: A reconsideration. *Journal of Economic Behavior & Organization*, 94:145–159, 2013.

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

[6] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk. Sparse factor analysis for learning and content analytics. *The Journal of Machine Learning Research*, 15(1):1959–2008, 2014.

[7] A. H. Murphy and R. L. Winkler. Scoring rules in probability assessment and evaluation. *Acta psychologica*, 34:273–286, 1970.

[8] J. Park and K. Williams. The effects of peer-and self-assessment on the assessors. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pages 249–254. ACM, 2016.

[9] A. Peysakhovich and M. Plagborg-Møller. Proper scoring rules and risk aversion. *Available at SSRN 2019078*, 2012.

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

[11] K. Raman and T. Joachims. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1037–1046. ACM, 2014.

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

[13] L. J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[14] R. Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1):43–62, 1998.

[15] R. Selten, A. Sadrieh, and K. Abbink. Money does not induce risk neutral behavior, but binary lotteries do even worse. *Theory and Decision*, 46(3):213–252, 1999.

[16] B. Strong, M. Davis, and V. Hawks. Self-grading in large general education classes: A case study. *College Teaching*, 52(2):52–57, 2004.

[17] K. Topping. Self and peer assessment in school and university: Reliability, validity and utility. In *Optimising new modes of assessment: In search of qualities and standards*, pages 55–87. Springer, 2003.

[18] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.