Measurement Integrity in Peer Prediction: A Peer Assessment Case Study

NOAH BURRELL and GRANT SCHOENEBECK, University of Michigan, USA

We propose measurement integrity, a property related to ex post reward fairness, as a novel desideratum for peer prediction mechanisms in many natural applications. Like robustness against strategic reporting, the property that has been the primary focus of the peer prediction literature, measurement integrity is an important consideration for understanding the practical performance of peer prediction mechanisms.

We perform computational experiments, both with an agent-based model and with real data, to empirically evaluate peer prediction mechanisms according to both of these important properties. Our evaluations simulate the application of peer prediction mechanisms to peer assessment—a setting in which ex post fairness concerns are particularly salient. We find that peer prediction mechanisms, as proposed in the literature, largely fail to demonstrate significant measurement integrity in our experiments. We also find that theoretical properties concerning robustness against strategic reporting are somewhat noisy predictors of empirical performance. Further, there is an apparent trade-off between our two dimensions of analysis. The best-performing mechanisms in terms of measurement integrity are highly susceptible to strategic reporting. Ultimately, however, we show that supplementing mechanisms with realistic parametric statistical models can, in some cases, improve performance along both dimensions of our analysis and result in mechanisms that strike the best balance between them.

CCS Concepts: • Theory of computation → Algorithmic game theory and mechanism design; • Computing methodologies → Agent / discrete models; • Social and professional topics → Student assessment.

Additional Key Words and Phrases: Measurement integrity, peer prediction, peer assessment

ACM Reference Format:
Noah Burrell and Grant Schoenebeck. 2023. Measurement Integrity in Peer Prediction: A Peer Assessment Case Study. In Proceedings of the 24th ACM Conference on Economics and Computation (EC ’23), July 9–12, 2023, London, United Kingdom. ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3580507.3597744

1 INTRODUCTION

Peer prediction [Miller et al., 2005], or information elicitation without verification, is a paradigm for designing mechanisms that elicit reports from a population of agents about questions or tasks in settings where ground truth (and thus the possibility of spot-checking) need not exist. An important dimension of evaluation for a peer prediction mechanism is the degree to which it rewards agents for their reports in a way that incentivizes truthfulness. This dimension, which we label as robustness against strategic reporting, has been the overwhelming focus of the theoretical...
peer prediction literature. We broaden this focus by introducing a new dimension of evaluation, measurement integrity.

In the peer prediction paradigm, we assume that agents receive a signal (perhaps at some cost) about each task, drawn from some joint prior distribution. In the current literature, mechanisms are typically characterized by two properties that attest to their robustness against strategic reporting:

1. An equilibrium concept related to truthfulness that the mechanism induces under certain assumptions.
2. The assumptions, which typically constrain the form of the joint prior distribution of signals for every agent, that are sufficient to ensure inducement of the equilibrium concept.

An appendix to the full version of the paper details these two properties for a representative selection of fundamental mechanisms from the peer prediction literature. However, these properties alone are insufficient for evaluating peer prediction mechanisms’ suitability for a given application. Firstly, this characterization omits other important desiderata. In many applications, for example, it is just as important for rewards to be fair as to be incentive-compatible. Secondly, even for a particular setting where incentive-compatibility is a primary desiderata for peer prediction mechanisms, this characterization fails to determine the best mechanism to use.

It is possible for a mechanism to induce a stronger equilibrium concept than another mechanism, but only under a stronger assumption. It is also possible for a mechanism to “approximately” induce a stronger equilibrium concept under a given assumption. In both cases, there is no clear answer to the question of which mechanism is more robust against strategic reporting. Considering secondary desiderata also discussed in the theoretical peer prediction literature does not help. Such properties, for example that mechanisms require little or no prior knowledge of the distribution of signals or that mechanisms only require simple reports from the agents, often fail to meaningfully differentiate the state-of-the-art mechanisms.

1.1 Our Contributions

- To address the insufficient characterization of peer prediction mechanisms, we introduce a new dimension of analysis. We call this new dimension measurement integrity and provide a formal definition alongside the motivating intuition.
- To address the issue of determining the best mechanism for a given application, we perform extensive computational experiments, using both synthetic data and real data, to evaluate empirical properties of peer prediction mechanisms in the context of an important purported application of peer prediction—peer assessment. First, we investigate the measurement integrity of state-of-the-art peer prediction mechanisms. Then, we broaden our experiments to develop a new, complimentary perspective on the robustness against strategic reporting. Together, the results of these experiments meaningfully differentiate the state-of-the-art peer prediction mechanisms in the literature (Figure 1). These experiments also serve as a guide for comparing peer prediction mechanisms in other settings.

1.2 Measurement Integrity

Fundamentally, measurement integrity quantifies the strength of a mathematical relationship between the quality of an agent’s reports and the reward they are allocated by a mechanism. Below, we define measurement integrity to be a concrete empirical estimand for which we can develop practical estimation strategies. The motivation for this definition, though, arises from its ability to represent a more abstract theoretical estimand: ex post fairness, where ex post, from the perspective of an agent, relates to all randomness from the mechanism’s choices subsequent to their making a decision (e.g., choosing their reporting strategy) or related to actions of the other agents.
In peer assessment, for example, \textit{ex post} fairness requires acknowledging that for a student, receiving an $A$ with 80\% probability and an $F$ with 20\% probability is not the same as certainly receiving a $B$; students should receive the grade they earn. Similarly, rewards should faithfully reflect the quality of the work submitted. We will see that mechanisms with high measurement integrity produce rewards that reliably reflect the quality of participants’ reports. In general, when participants reflect on their experience with a mechanism and assess the fairness of that interaction, we believe they will tend to ask fundamentally \textit{ex post} questions like “Was my reward fair compensation for my effort?” rather than \textit{ex ante} ones. Measurement integrity speaks directly to these kinds of questions. As a result, taking measurement integrity seriously is an important step in transforming peer prediction mechanisms from intellectual curiosities into practical tools for eliciting information in the real world.

Moreover, the relationship between agent qualities and rewards at the heart of measurement integrity, and \textit{ex post} fairness more generally, is useful for other goals: The rewards of a mechanism with high measurement integrity identify agents with high-quality reports—an important component of many strategies to aggregate information elicited from a population of agents with different levels of proficiency at the given task. On the other hand, a mechanism with low measurement integrity will have a noisy relationship between agent quality and rewards, which in tournament settings has been shown to increase the optimal payment required to elicit a certain effort level [Drugov and Ryvkin, 2020].

1.2.1 Defining Measurement Integrity. Suppose that $P$ is a data-generating process (DGP) for a given application (our model is described formally in Section 2), which, along with a (deterministic) quality function $Q$ and (stochastic) mechanism $M$, produces (1) a vector of agent report qualities $q$ and (2) a vector of agent rewards $r^M$. These vectors have dimension $n$, the number of agents from $P$.

Given this, the last component of the definition of measurement integrity is a correlation function $\text{corr} : \mathbb{R}^n \times \mathbb{R}^n \to [-1, 1]$. Correlation functions, also called correlation coefficients, describe the strength of a mathematical relationship between two quantities. For us, these two quantities are an agent’s reward and an agent’s quality. The absolute value of the correlation function increases with the strength of the relationship from 0, indicating no relationship, to 1, indicating a perfect correspondence. The sign of a non-zero correlation indicates whether changes in one variable are associated with the same kind of changes in the other variable—a positive relationship—or with opposite kinds of changes—a negative relationship.

\textbf{Definition (Measurement Integrity).} The \textit{measurement integrity} of a peer prediction mechanism $M$ with respect to a data-generating process $P$, a quality function $Q$, and a correlation function $\text{corr}$ is

$$MI_{P,Q,\text{corr}}(M) = \mathbb{E}_{P,M} \left[ \text{corr} \left( q, r^M \right) \right].$$

1.2.2 Unpacking the Definition. One crucial component of the definition of measurement integrity is the correlation function. Correlation functions are not an arbitrary class; they have several key features that make them uniquely appropriate. First, they require reference points for perfect, perfectly opposite, and non-existent relationships. This lends the values of correlation functions a relative interpretability that is absent from functions whose values have more absolute meaning, e.g., generic loss functions, which only require a reference point for perfect performance (0).

Intuitively, correlation functions quantify the extent to which rewards reflect the precise features of the qualities that are most important for a particular application. This is important, because practitioners in different contexts may care about different kinds of such relationships. For example, a practitioner might want (1) rewards to be proportional to some notion of quality (e.g., the inverse
of absolute error); (2) the order of the rewards to represent the order of the reports’ quality; or (3) the top 50% of agents recognized as such. Different (classes of) correlation functions, which measure the strength of different kinds of relationships, allow us to capture each of these different contexts, because correlation functions can be chosen so that their values are invariant under exactly the set of transformations (applied to the arguments) that preserve the relationship of interest. In the above examples, (1) Pearson correlation; (2) Kendall rank correlation; and (3) a transformation of area under the ROC curve are appropriate choices; we will return to these examples in our experiments.

Moreover, many mechanisms use peer prediction as an intermediate step in assigning rewards. Peer prediction mechanisms output scores that typically must be transformed into rewards in a manner that takes into account the particular context (cost, effort, etc.) and different mechanisms use different transformations. For example, scores may be linearly scaled or used to reward agents based on their rank or quantile in the agent population. In such cases, the rewards are unchanged by certain transformations of the scores (e.g., linear or monotone). Thus, to determine whether a peer prediction mechanism outputs useful scores in a given context, it is best to use an evaluation metric that is similarly invariant. Correlation functions, which exemplify such evaluation metrics, are thus more suitable than generic loss functions, which require raw scores to be directly comparable.

The idea of invariance to a set of transformations also establishes the connection between correlation functions and measurement scales that inspires the term measurement integrity. Stevens [1946], in a seminal work on measurement, characterizes four scales in terms of the set of transformations under which the information communicated by a particular measurement is invariant.\(^1\)

Another crucial component of the definition of measurement integrity is that it goes beyond single evaluations of correlation functions. It considers the value of the correlation function across all of the potential outcomes of the data-generating process and the given mechanism. This is important, because it establishes that, for a mechanism with high measurement integrity, the relationship between rewards and qualities must be consistently strong.\(^2\) The need for this consistency is at the heart of why measurement integrity is important in many peer prediction applications.

Lastly, the dependence of measurement integrity on a particular data-generating process is crucial. In this way, measurement integrity is similar to many more familiar quantities like precision, recall, accuracy, and other evaluation metrics from machine learning. Such metrics similarly have values that depend on the particular data-generating process under which they are evaluated, but contextualize those values in a consistent, useful way. In the measurement theory literature, these are called relative scales and allow the creation of measurements on an absolute scale to enable us to “capture an underlying order in a complex problem” even though “such scales cannot exist objectively for all time and all objects, but only for a certain time and a given set of objects.” [Saaty, 2004]. This provides another contrast with generic loss functions, which, while absolute, are more difficult to compare across different contexts (e.g., when the number of agents is different).

### 1.3 Our Approach

Alongside the relatively abstract definition, it is useful to have an explicit strategy for computing useful estimates of realistic measurement integrity values for a given application. Our work that follows is largely devoted to demonstrating such a strategy. We apply a similar strategy for robustness against strategic reporting in the full version of the paper. The heart of our approach is to focus on empirical quantities, rather than analytically-derived ones. We estimate measurement

---

1. To highlight this connection, we reference specific scales, e.g., in the term ordinal measurement integrity, when the value of the relevant correlation function is invariant under the same transformations as that scale, e.g., ordinal, applied to \(r^M\) or \(q\).

2. It is worth noting that expectations in general do not have this property—they can be driven up by extremely high values that occur with relatively low probability. However, for the expectation in the definition of measurement integrity, this is not the case, because correlation functions are bounded above by 1.
integrity in computational experiments, first with a realistic agent-based model (ABM) and then again with real data. Computational experiments offer important advantages over analytical or theoretical approaches for analyzing measurement integrity. Computational experiments are more naturally outcome-oriented. Simulated outcomes can be generated cheaply, frequently, and reliably under a wide range of parameter specifications. In contrast, making general theoretical statements quantifying *ex post* payments from a mechanism in this setting is cumbersome and difficult. While theorems have the advantage of potentially applying to a larger range of settings, we believe that, in this context, theorems are unlikely to give tight bounds and likely to be hard to interpret. In contrast, our computational experiments readily provide interpretable results, albeit on a chosen set of inputs. Further, our definition of measurement integrity requires fixing a specific data-generating process. This inherently limits the generality of any particular estimate of measurement integrity and, consequently, limits the potential benefits from the generality of a theoretical approach. We will see firsthand why this is essential, as the peculiarities of the setting we choose to explore—peer assessment—appear to be an important driver of our results in this work. We will also see, however, that this is not too limiting. Our results qualitatively accord across the related peer assessment data-generating processes—simulated and real—that we consider.

1.3.1 Peer Assessment. As emphasized above, our definition of measurement integrity is always instantiated with respect to a particular data-generating process. Thus, the usefulness of our estimates of measurement integrity for different mechanisms depends on the fidelity of the data-generating process or processes under which those estimates are derived to the real application of interest. With this in mind, we believe that peer assessment—where students provide feedback on the work of their peers—is a good candidate application for which to apply our techniques.

First, realistic models for data-generating processes (Section 5) in peer assessment exist and real peer assessment data (Section 2.2) are available for us to use in our computational experiments.

Second, peer assessment involves a common purported application of peer prediction: peer *meta-grading*. Meta-grading is the task of “grading the graders”: assessing the quality of peer assessments. In peer grading, the primary challenge is to correctly aggregate peer reports into grades that reflect the quality of a submission for an assignment. In peer meta-grading, incentive concerns are more salient. A student’s grade for an assignment depends on other students’ feedback (and the quality of their own submission). A student’s meta-grade depends directly on their own feedback for other students. As a result, mechanisms for which the incentives are poorly designed will discourage rather than encourage high-quality feedback.

Lastly, meta-grading presents some unique challenges for the peer prediction paradigm that make it a particularly interesting case study:

*Fairness Constraints*. Grades are intended to reflect the quality of students’ work, so the meta-grades assigned by any peer assessment mechanism must reflect the actual quality—not just the expected quality—of each student’s performance in their assigned peer assessment tasks. To be suitable for assigning meta-grades, then, peer assessment mechanisms, should demonstrate significant measurement integrity.

*Heterogeneous Quality*. Agents may have different skills, exert different effort levels, and may not be fully calibrated. Mechanisms may vary in their ability to handle such differences.

*Scarcity of Data*. Which mechanism performs best may be sensitive to the amount of data available. In peer assessment, the amount of data is severely constrained. There are limits on the number of assignments students can be expected to complete in a course (often around 10) and the number of submissions students can be expected to grade per assignment (typically between 3 and 6).
To complete our formal definition of measurement integrity, we first specify the necessary components of a data-generating process. At a high level, $P \rightarrow (I, J, G)$ generates a set of agents $I$, a set of tasks $J$, and an assignment graph $G$ of agents to tasks. Each of these components may be characterized by a set of parameters described by $P$. More specifically, $P$ should, at minimum: (1) generate exactly one ground truth response for each task $j$ and (2) generate a signal $s(i, j)$ that agent $i$ perceives about the ground truth for task $j$ for each edge $(i, j) \in G$. Further, $P$ may also describe how each agent $i$ computes a report to submit to a mechanism for each edge $(i, j) \in G$.

3Measurement integrity is computed wrt the ABM and real data DGPs if the number of assignments/assignment blocks were uniformly random, MSE from ground truth, and the Kendall rank correlation coefficient. Empirical robustness against strategic reporting is computed wrt the analogous DGPs in if the number of strategic graders and their strategy were uniformly random among those we consider and mean rank gain. Unfamiliar terms in these specifications are defined in Section 4 and in the full paper, respectively.

Fig. 1. Direct, two-dimensional comparisons of peer prediction mechanisms that broadly summarize our experimental results with ABM and real data. For the real data, the final point values are computed by taking the average results over all four semesters in the dataset (Section 2.2), whereas uncertainty is estimated using the maximum and minimum values of the relevant quantities across the individual semesters.

Mechanisms are colored according to their theoretical robustness against strategic reporting, as described by the relevant equilibrium concept (see full version). When compared to the x-axis, it is clear that theoretical robustness is a somewhat noisy predictor of empirical robustness in our experiments.

1.4 Our Results

Our experimental results, summarized in Figure 1, robustly differentiate existing peer prediction mechanisms according to their empirical performance. Moreover, they indicate an apparent trade-off inherent in seeking to simultaneously optimize measurement integrity and empirical or theoretical robustness against strategic reporting. They also reveal the following lessons:

- Generic peer prediction mechanisms from the literature largely fail to demonstrate significant measurement integrity compared with simple baselines.
- Measurement integrity and empirical robustness against strategic reporting create a more fine-grained comparison between mechanisms than theoretical notions of robustness against strategic reporting alone, since those notions are often not directly comparable.
- Certain mechanisms can be augmented with parametric statistical models to improve their measurement integrity and empirical robustness against strategic reporting. Consequently, parametric mechanisms should receive more attention in the peer prediction literature.

2 DATA-GENERATING PROCESSES

To complete our formal definition of measurement integrity, we first specify the necessary components of a data-generating process. At a high level, $P \rightarrow (I, J, G)$ generates a set of agents $I$, a set of tasks $J$, and an assignment graph $G$ of agents to tasks. Each of these components may be characterized by a set of parameters described by $P$. More specifically, $P$ should, at minimum: (1) generate exactly one ground truth response for each task $j$ and (2) generate a signal $s(i, j)$ that agent $i$ perceives about the ground truth for task $j$ for each edge $(i, j) \in G$. Further, $P$ may also describe how each agent $i$ computes a report to submit to a mechanism for each edge $(i, j) \in G$.

3Measurement integrity is computed wrt the ABM and real data DGPs if the number of assignments/assignment blocks were uniformly random, MSE from ground truth, and the Kendall rank correlation coefficient. Empirical robustness against strategic reporting is computed wrt the analogous DGPs in if the number of strategic graders and their strategy were uniformly random among those we consider and mean rank gain. Unfamiliar terms in these specifications are defined in Section 4 and in the full paper, respectively.
as a function of their signal \( s(i, j) \). Without this specification, we assume that an agent’s report is always truthful, i.e., equal to their signal.

Now, suppose a population of agents \( I \) and of tasks \( J \) has been drawn, the tasks have been assigned to the agents, and corresponding reports have been generated, all according to \( P \). The reports are then submitted to a peer prediction mechanism \( M \) that computes a vector of rewards \( r^M = M(I, J, G) \). For each agent \( i \in I \), their individual reward is \( r^M_i \), the \( i \)-th component of \( r^M \). Corresponding to the vector of rewards is a vector of qualities \( q = Q(I, J, G) \), where \( Q \) is referred to as a quality function. Unlike mechanisms, we note that quality functions must be deterministic given a complete instance generated by \( P \). For each agent \( i \in I \), the \( i \)-th component of \( q \), \( q_i \), quantifies the quality of \( i \)'s reports. For example, \( q_i \) may be the mean squared error of \( i \)'s reports with respect to known ground truth values. We now turn to the specific peer assessment data-generating processes that we consider in this work.

### 2.1 Peer Assessment Agent-Based Model

Our ABM simulates a class of students enrolled in a semester-long course for which there is at least one graded assignment. For each assignment, each student turns in one submission and is randomly assigned submissions from four other students to grade.

**Submissions.** For each assignment \( j \), each student \( i \) turns in a submission \( s_{i,j} \). Each submission \( s_{i,j} \) has a true integer score \( g^*_{i,j} \in [0, 10] \), drawn (independently at random) from the binomial distribution \( B(10, \frac{7}{10}) \).

**Signals.** The process of grading is modeled by an agent receiving a signal about the true score of a submission. The signal is a function of some number of draws from a latent distribution that depends on the true score of the submission being graded and the bias of the agent.

**Bias.** In practice, agents may have some bias in grading assignments. That is, the latent distribution from which draws are used to construct their signal for each assignment could have a mean that is slightly higher or lower than that assignment’s ground truth score. An agent \( k \)'s bias \( b_k \) is sampled uniformly at random from the normal distribution \( N(0, 1) \). Their signal for submission \( s_{i,j} \) is a function of draws from the latent distribution \( B \left( 10, \frac{g^*_{i,j} + b_k}{10} \right) \). If \( g^*_{i,j} + b_k \) is less than 0 or greater than 10, then the value is truncated to be 0 or 10, respectively.

**Draws from Latent Distribution.** The number of draws from an agent’s latent distribution that are used to create their signal—which determines the variance of the distribution of their signals—is a function of their effort; greater effort corresponds to lower variance.

When the signal is created using a single draw, defining the signal is trivial—the signal is equal to the outcome. When the signal is created using more than one draw from the latent distribution, the signal is defined as the simple average of the outcomes of the draws rounded to the nearest integer. This convention ensures that the space of signals is equal to the space of reports, so the notion of a “truthful report” is straightforward and well-defined. Our model of effort is as follows:

**Continuous Effort.** Effort is parameterized by a continuous value \( \lambda \in (0, 2] \) drawn uniformly at random. The number of draws from the latent distribution used to create an agent’s signal is equal to \( 1 + X \), where \( X \sim \text{Pois}(\lambda) \) is drawn according to the Poisson distribution.

**Report Quality.** We use a simple, intuitive notion of report quality—the squared distance between the report value and the true grade of the corresponding submission. In the full version of the paper, we discuss alternative conceptions of report quality and their relative merits in detail.
2.2 Peer Assessment Data

Our real peer grading data set—which was collected for other projects [Yuan and Downey, 2020] and graciously shared with us for this work—contains grading information from an undergraduate-level course on the design and analysis of algorithms taught at Northwestern University in both the Spring and Fall semesters of 2017 and 2019. For each student enrolled in the course, the data set contains information about the submissions they turned in during the course of the semester and the grades that they provided for submissions from other students. For each submission, the data set identifies the assignment that the submission corresponds to, specifies the grade that was ultimately awarded for that submission—which we treat as its true grade—and some number of peer grades. These awarded grades are a mix of grades assigned by instructors and grades assigned by the (non-parametric) vancouver algorithm [de Alfaro and Shavlovsky, 2014], which takes into account instructor grades, peer grades, and the accuracy of peer graders. Because this combination of methods was deemed sufficient for fairly assigning grades to real students in the courses from which the data were collected, we are comfortable treating the assigned grades as unbiased estimates of the ground truth. For each peer grade that a student provided, the data set includes an identifier for the corresponding assignment, a numerical score, and written comments.

Many of the peer prediction mechanisms that we consider impose restrictions on the form of the data set. Certain mechanisms require that students grade at least two submissions for each assignment in which they will be evaluated by the mechanism. Other mechanisms require that at least two students grade each submission. Accordingly, we (iteratively) remove peer grades from students for assignments for which they graded fewer than 2 submissions and submissions with fewer than 2 graders until the modified data set meets these specifications. After this pre-processing, we simplify the grading and report space. The numerical scores—true grades and peer grades—from 2017 are out of 100 and from 2019 are out of 30. We coarsen these raw grades into the integer range [0, 10]. This simplifies the implementation of the peer prediction mechanisms and our experiments by keeping the space of possible reports the same for our ABM and all 4 semesters in the data set and helps to make the empirical distribution of reports less sparse in the space of all possible reports. It also lends our analysis some robustness to the method of assigning true grades, since small changes to the value of a true grade will tend not to change the value to which it is mapped during the pre-processing procedure.

Lastly, our parametric mechanisms (Section 3.3) (and, in the full version of the paper, certain reporting strategies) employ information about a (continuous) prior distribution for the true grades. Thus, we use maximum likelihood estimation to fit normal distributions to the empirical distribution of true grades for each semester.

The relevant information about each semester after all of the pre-processing is given in Table 1.

3 PEER PREDICTION MECHANISMS

In this work, we consider a representative selection of fundamental mechanisms from the peer prediction literature. In what follows, we describe the intuition behind the mechanisms that we evaluate using our agent-based modeling framework. We also discuss the challenges that the particularities of the peer assessment setting and our peer assessment model pose to the implementation of the mechanisms. For a more specific discussion of the actual implementation of the various mechanisms and how we overcome these challenges, see the full version of the paper.

3.1 Baseline

In theoretical work, simple baselines would typically be excluded, due to the existence of trivial, non-truthful reporting equilibria. Despite this concern, however, such simple mechanisms are used
### Table 1. Summary of the pre-processed peer grading data, including the total number of students, assignments, submissions, and peer grades for each semester. The percentage of grades in the raw data that are retained after pre-processing is also shown. For example, in Spring 2019, the 1586 grades left after pre-processing represent 91.9% of the total grades in the raw data. Lastly, the parameters $\mu$ and $\sigma^2$ are the mean and variance, respectively, of the normal prior distribution fit to the empirical distribution of true grades. These values are used in the implementations of parametric peer prediction mechanisms.

| Semester   | Students | Assignments | Submissions | Peer Grades | Retained  | $\mu$  | $\sigma^2$ |
|------------|----------|-------------|-------------|-------------|-----------|--------|-----------|
| Spring 2017| 94       | 16          | 758         | 4080        | 99.9%     | 8.71   | 3.8025    |
| Fall 2017  | 86       | 16          | 577         | 3102        | 99.9%     | 7.57   | 4.9729    |
| Spring 2019| 49       | 13          | 313         | 1586        | 91.9%     | 7.68   | 3.6864    |
| Fall 2019  | 65       | 14          | 389         | 2089        | 98.1%     | 8.25   | 2.8561    |

in practice. Bachelet et al. [2015], for example, recommend using the following mechanism to assign grades (when the reports have been appropriately pre-processed):

**Mean Squared Error (MSE) Mechanism.** On each submission that they grade, agents are paid according to the mean squared error of their reports from the consensus grade of each submission. The consensus grade of a submission—a basic estimate of its unobservable true grade—is defined to be the simple average of the reports of all 4 agents that graded it. To maintain the convention that the higher rewards correspond to higher quality agents, the payments are equal to the negative of the mean squared error.

### 3.2 Non-Parametric Mechanisms

Our first category of peer prediction mechanisms reflects the options that a novice mechanism designer would find in an initial search for peer prediction mechanisms to deploy in some application. In keeping with this, we implement these mechanisms as faithfully as possible to the descriptions given in the works in which they were proposed. We make changes only when necessary to ensure basic functionality within the setting of our model.

Note that for all mechanisms that involve pairing an agent with another agent in order to compute their scores on a grading task (i.e., generating a report for one submission), we take the expectation over all of the possible pairings to reduce the variance of the scores.

**Output Agreement (OA) Mechanism.** The simplest type of peer prediction mechanism, common in the literature [Faltings et al., 2017a], is an output agreement mechanism, which serves as another simple baseline with which to compare state-of-the-art peer prediction mechanisms. To compute payments for a task in the OA mechanism, agents are paired and their reports are compared. Agents are paid 1 if their reports match and 0 otherwise.

**Peer Truth Serum (PTS) Mechanism.** Developed by Faltings et al. [2017a], the PTS mechanism pays agents if their report for a task agrees with the report of a randomly selected peer on the same task. The magnitude of the payment is proportional to the inverse of the frequency of their report according to a distribution $R$ over the report space.

**$\Phi$-Divergence Pairing ($\Phi$-Div) Mechanism.** This mechanism was proposed by Schoenebeck and Yu [2021] and is based on the application of an information-theoretic framework for designing peer prediction mechanisms described by Kong and Schoenebeck [2019]. Like the OA mechanism, this mechanism pairs agents with peers. The pairs are rewarded for submitting correlated reports on a
**bonus task** and penalized for submitting correlated reports on a pair (one for each agent) of **penalty tasks** that are distinct from each other and from the bonus tasks.

The magnitudes of the respective reward and penalty depend on a convex function \( \Phi \) chosen by the mechanism designer and on \( JP(x, y) \), the joint-to-marginal-product ratio of random variables \( X \) and \( Y \) drawn, respectively, from each agent’s distribution of reports:

\[
JP(x, y) = \frac{P_{X,Y}(x, y)}{P_X(x)P_Y(y)},
\]

where \( P_{X,Y}(x, y) \) is the probability of observing reports \( x \) and \( y \) as answers to the same question under the joint distribution of reports \( X \) and \( Y \), and \( P_X(x)P_Y(y) \) is that probability according to the product of the marginal report distributions. Their ratio can be understood as measuring how much likelier a pair of reports \( x \) and \( y \) is to occur on the same question versus different questions. Note that each quantity is a function of the random variables \( X \) and \( Y \), which depend both on the agents’ strategies and the joint prior. In general, \( JP \) is unknown and will need to be estimated.

For a given pair of agents, a **bonus task** \( b \), and a pair of **penalty tasks**, \( p \neq q \), the payment is:

\[
\partial \Phi(JP(x_b, y_b)) - \Phi^*(\partial \Phi(JP(x_p, y_q))),
\]

where \( \partial \Phi \) is the subgradient of \( \Phi \), \( \Phi^* \) denotes the convex conjugate of \( \Phi \), and \( x_i \) and \( y_j \) denote the first agent’s report on task \( i \) and the second agent’s report on task \( j \), respectively. Intuitively, the first term rewards an agent based on the likelihood of their report for the bonus question \( b \) given the other agent’s report on \( b \). The second term penalizes agents for reporting generically likely answers by considering the likelihood of an agent’s report for the penalty question \( p \) given the other agent’s report on the distinct penalty question \( q \). See Schoenebeck and Yu [2021] for a complete discussion of each component of this mechanism, including definitions of the relevant terms above.

Ideally, we would want to estimate the joint-to-marginal-product ratio of the reports for each pair of agents, but given the limited availability of data in this setting, the best we can do is treat the agents anonymously and compute one estimate, \( \hat{JP} \), that applies to the entire agent population. See the full version for the details of this estimation procedure.

In our experiments, we consider four common \( \Phi \)-divergences, described in Table 2. One important note with respect to the choice of \( \Phi \)-divergence is that when total variation distance (TVD) is chosen, the \( \Phi \)-Div mechanism emulates the Correlated Agreement mechanism from Shnayder et al. [2016a] (which in turn generalizes the ideas underlying the mechanism proposed by Dasgupta and Ghosh [2013]).

One significant omission from our selection of state-of-the-art mechanisms is the Determinant-based Mutual Information (DMI) mechanism [Kong, 2020], which has impressive theoretical properties. Without significant modifications to the setting of our experiments, it will simply assign every agent a reward of 0. With the modifications, the mechanism does not demonstrate high measurement integrity and the results for robustness against strategic reporting are not a fair comparison to the other mechanisms. As a result, we omit DMI from consideration in this work. However, we discuss our implementation and the necessary modifications in the full version of the paper and include DMI, when possible, in the additional experimental results in that version.

### 3.3 Parametric Mechanisms

Anticipating the challenges that generic mechanisms might encounter when deployed in a specific setting, we also explore how certain peer prediction mechanisms can be supplemented with domain-specific, parametric statistical models. To implement these, we adopt the perspective of a real-life mechanism designer. In the real world the “true” distributions and parameter values that “govern” the behavior of students participating in peer assessment are inaccessible. Instead, a mechanism...
designer can examine the peer assessment literature to find a model of peer assessment inspired by and validated on real data for which the hyperparameters of the model can be tuned to fit their particular application. Here, model PG₁ from Piech et al. [2013] meets both criteria. It constitutes a reasonable continuous approximation to our primarily discrete underlying model (in which “reliability” serves as a proxy for effort). The model, with hyperparameters that are appropriate for our setting, is described below:

- **True Score**: \( g_{i,j}^* \sim \mathcal{N}(7, 2.1) \) for each submission \( s_{i,j} \),
- **Reliability**: \( \tau_i \sim \mathcal{G}(10/1.05, 10) \) for each agent \( i \),
- **Bias**: \( b_i \sim \mathcal{N}(0, 1) \) for each agent \( i \),
- **Signal**: \( z_{i,j}^k \sim \mathcal{N}(g_{i,j}^* + b_k, \tau_{i}^{-1}) \) for a grader \( k \), who is grading submission \( s_{i,j} \).

To reiterate, the simulated data in our experiments is always generated according to the model described previously in Section 2.1. But instead of estimating parameters of that underlying model, we estimate the parameters of model PG₁. We then use those estimates in deploying the parametric peer prediction mechanisms described below. This simulates the situation faced by a mechanism designer in a real deployment. They would be unable to know the “true” underlying model, but would be able to tune a reasonable statistical model for their application using past data. Using model PG₁ is also useful because existing work from the peer assessment literature shows how to estimate its parameters. Chakraborty et al. [2020] propose a method for estimating the parameters of model PG₁ (and computing meta-grades) using limited access to ground truth. In the absence of ground truth, their estimation method (though not their meta-grading method) can be adapted to estimate the parameters of the model using an expectation-maximization-style algorithm with Bayesian priors for the bias and reliability of each agent. The details of our estimation procedure are available in the full version of the paper.

**Parametric MSE (MSE_p) Mechanism.** Under this mechanism, each agent is awarded according to the mean squared error of their reports (corrected for estimated biases) from the estimated true scores. As with the baseline, the payments are equal to the negative of the mean squared error.

**Parametric Φ-Divergence Pairing (Φ-Div_p) Mechanism.** Instead of using empirical estimates of the joint-to-marginal-product ratio of reports, we can pre-compute the joint-to-marginal-product ratio JPh(x, y) analytically under model PG₁ (see full version) and score the tasks after estimating the parameters of model PG₁ using the estimation procedure described above. This allows us to individualize the joint-to-marginal-product ratio for each pair of agents, which is desirable but intractable for the non-parametric version of this mechanism, given the scarcity of data.

\[^4\]G denotes a Gamma distribution. The hyperparameters \( a_0 = 10/1.05 \) and \( b_0 = 10 \) for \( G \) were chosen by inspection, subject to having the correct expected value for a continuous effort agent.
For each task, agents are paired and scored via the same procedure as the non-parametric \( \Phi \text{-Div} \) mechanism, but using the expression we derived for \( JP(x, y) \) instead of an empirical estimate.

4 QUANTIFYING MEASUREMENT INTEGRITY

Overall, we seek to empirically evaluate the above mechanisms according to both their measurement integrity and robustness against strategic reporting. Evaluating mechanisms for both properties simultaneously would make it difficult to isolate which features were beneficial for which property. As a result, in this paper, we quantify measurement integrity in isolation, assuming that agents report their signals honestly. In the full version of the paper, we conduct further experiments to also quantify empirical robustness against strategic reporting.

4.1 Computational Experiments with ABM

One key advantage of the agent-based modeling approach is access to latent quantities, e.g., a submission’s true score, that are generally not observable, without noise, in the real world. Another is the ability to readily repeat experiments over a range of parameter specifications. We leverage these advantages in order to analyze the relationship between agents’ payments assigned under the various mechanisms and the squared error of their reports to the ground truth scores, considering various intuitive notions of measurement integrity.

4.1.1 Methods. For each mechanism, we perform the same procedure of simulating “semesters,” which consist of 500 students submitting and grading some number of simulated assignments. The number of assignments is varied from \( i = 1, 2, \ldots, 15 \). For every value of \( i \), we simulate 50 semesters, which each proceed as follows:

(1) For each assignment:
   i. All students turn in a submission characterized by a true grade.
   ii. A random 4-regular graph of agents is constructed.
   iii. Students grade the submissions of their neighbors in the graph according to our peer assessment model. The squared errors of their grades to the true grades are recorded.
   iv. The grades are reported to the mechanism, which assigns a reward to each student for their performance in peer assessment for that assignment.

(2) Correlation functions are evaluated using students’ cumulative rewards—the sums of their reward for each individual assignment—and their cumulative squared error to the true scores.

4.1.2 Correlation Functions. Choosing an appropriate correlation function is crucial to operationalizing measurement integrity as a useful quantity for a peer prediction application. There are already many useful correlation coefficients that correspond well to the various measurement scales. In addition to these, measurement integrity can also accommodate more customized correlation functions. This is useful, because the best correlation function for a given setting can depend on a mechanism designer’s tolerance for making different kinds of errors in that application. In peer grading, for example, a practitioner may consider it worse to fail a borderline student that should have passed than to pass a borderline student that should have failed. Their choice of correlation function, then, may take these preferences into account.

To illustrate these principles, in our experiments, we consider increasingly fine-grained measurements. For the first of these, we illustrate how we can modify a popular machine learning metric that is well-suited to evaluate that kind of measurement into a new correlation function. Afterward, we adopt well-known correlation functions that are similarly well-suited for their corresponding kinds of measurement. In doing this, we note that we depart somewhat from the approach that we would expect others to take in applying our techniques. Generally, we would expect that a mechanism
designer will have a utility function in mind that will help them determine a particular correlation function. We take a more exploratory approach—considering multiple correlation functions without a particular utility function in mind—in order to find out which kinds of measurement integrity, if any, are high for the peer prediction mechanisms we consider in our case-study application.

**Coarse Ordinal Measurement.** First, we consider binary classification of agents as being above or below the median in terms of squared error to ground truth scores. For binary classification tasks, the area-under-the-curve (AUC) of the receiver operating characteristic (ROC) curve is an evaluation metric that is widely used in machine learning. This metric is particularly well-suited to evaluating mechanism performance at our binary classification task: AUC summarizes how useful the rewards assigned by the mechanism are for being translated into a classifier via the selection of a threshold such that all students with rewards above the threshold are classified as above-median and all students with rewards below the threshold are classified as below-median. At first glance, AUC does not quite meet our definition of a correlation function, because it varies between 0 and 1, not -1 and 1. However, it does have values that impart analogous meanings to those required by a correlation function: An AUC score of 1 indicates a perfect classifier, an AUC score of 0.5 is the expected value of a random classifier, and an AUC score of 0 indicates a perfectly opposite classifier. Thus, we can transform AUC into a new correlation function—AUC correlation (AUCC)—by rescaling the significant values:

\[
\text{AUCC} = 2 \cdot \text{AUC} - 1.
\]

**Fine Ordinal Measurement.** The most fine-grained ordinal measurement is ranking. For rankings, useful correlation functions already exist. Here, we adopt the Kendall rank correlation coefficient (\(\tau_B\)). The value of \(\tau_B\) is related to the number of pairs that appear in the same order (concordant pairs) and in the opposite order (discordant pairs) in the two rankings being compared. In the case that neither ranking has ties, \(\tau_B\) is equal to the proportion of pairs that are concordant minus the proportion of pairs that are discordant.

**Interval Measurement.** It is conceivable that the rewards from some mechanisms might contain even more fine-grained information about agents’ squared error from the ground truth than just the ordinal notion that higher payments correspond to lower squared error. That is, the magnitude of the difference in payments between agents may contain information about the magnitude of their difference in squared error. At least, this should be the case for the baseline MSE and MSE<sub>P</sub> mechanisms, if they are computing good estimates of the unobserved (by the mechanisms) ground truth scores. In particular, we would expect that those mechanisms would contain good linear information about agents’ squared error. That is, a given magnitude of the difference in reward should more or less correspond to the same magnitude of difference in quality regardless of the particular reward values. This property is characteristic of interval measurement, and can be evaluated with the most familiar correlation coefficient: the Pearson correlation coefficient (\(\rho\)), which measures the strength of the linear relationship between two variables.

4.1.3 Results. Estimates of measurement integrity with respect to the various correlation functions, while varying the number of assignments per semester, are plotted in Figure 2. These estimates are computed by taking an average of the value of the correlation function over the 50 semesters simulated for each value of the number of assignments in a semester.

The first result that stands out is that it is feasible to achieve high levels of measurement integrity, even under strict measurement scales. As the number of assignments in a semester increases, thereby increasing the amount of information available to the mechanisms, the baseline MSE and MSE<sub>P</sub> mechanisms score consistently highly according to each of the correlation functions,
including near-perfect Pearson correlation. This indicates that, unsurprisingly, it is possible to estimate true scores that are not observed by the mechanism highly reliably in our model when agents report truthfully. However, despite this possibility, peer prediction mechanisms generally do not appear to take advantage of this. As a result, they largely perform relatively poorly according to each correlation function compared to simple baseline mechanisms. The exceptions are two parametric mechanisms, $\Phi$-Div$_P$: KL and $\Phi$-Div$_P$: $H^2$, which mostly outperform the OA baseline.

Interestingly, the pattern established in the plots of the first two correlation functions, which are qualitatively nearly identical, does not hold in the plot of the Pearson correlation ($\rho$). In particular, the $\Phi$-Div$_P$: $H^2$ mechanism performs much less well according to Pearson correlation than the other correlation functions, indicating high ordinal measurement integrity, but relatively low interval measurement integrity. Importantly, this shows that the measurement scale that is relevant to a particular application—the notion of measurement that is desirable—can matter significantly with respect to how a mechanism performs. In this case, inspection reveals that the difference in performance is due to a tendency of the $\Phi$-Div$_P$: $H^2$ mechanism to occasionally assign very negative outlier payments. These outliers interfere with the linear relationship between the payments and agents’ squared error, but not the ordinal relationship. The payments from the PTS mechanism are also notably less useful with respect to linear than ordinal information. However, the ordinal information conveyed via the PTS mechanism was already poor relative to the other mechanisms.

### 4.2 Computational Experiments with Real Data

#### 4.2.1 Methods

We replicate the experiments with simulated data described in Section 4.1.1, substituting our four semesters worth of real data for the 50 semesters worth of data that we simulated via our peer grading ABM. For each semester, as in our experiments with our ABM, we assign rewards according to each mechanism for each assignment, one at a time. However, due to limitations in the data and the necessary pre-processing (Section 2.2), not every student is
associated with peer grades for submissions on every assignment. In fact, different students are
associated with different numbers of peer grades, which complicates our analysis.

We address this complication in two steps. First, we divide each student’s squared error of reports
from true scores and their payments by the number of peer grades with which they are associated.
Thus, we consider the mean squared error and average payment of each student when evaluating
the correlation functions. Second, to control for the fact that the mean squared error and payment
for students associated with few peer grades may be much noisier than those of students associated
with many peer grades, we focus on students associated with many peer grades when computing the
values of the correlation functions. To accomplish this, for each semester, we split the assignments
into four blocks of roughly equal size and consider only students who are associated with at least
one peer grade in each assignment block when evaluating the correlation functions. Under this
rule, the number of students considered when evaluating the correlation functions is 84, 49, 42, and
54 for the Spring 2017, Fall 2017, Spring 2019, and Fall 2019 semesters, respectively. Overall, the
correlation functions are evaluated 4 times for each semester—one after each assignment block
has been processed. Note that information accumulates as each block is processed—the evaluation
of the correlation function uses all the information obtained in the current block of assignments
and its predecessors. Thus, a nice consequence of this procedure is that new information for every
student is incorporated every time the correlation functions are re-evaluated.

Although there is more uncertainty in our estimates of measurement integrity in this setting,
due to our inability to re-draw samples from the underlying data-generating process, we are still
able to take the randomness of the mechanisms into account: We record the average value of the
correlation functions over 50 iterations of the procedure described above.

4.2.2 Results. We focus on the results for fine ordinal measurement—shown in Figure 3—for which
the correlation function is the Kendall rank correlation coefficient ($\tau_b$), since, as in our simulated
experiments, those results are qualitatively similar to those for coarse ordinal measurement and
since $\tau_b$ connects with our experiments for quantifying robustness against strategic reporting (see
full version), which involve rankings. The results for the other correlation functions are also given
in the full version of the paper.

Unsurprisingly, the results for the real data are much noisier than those for the simulated
data. However, there are some patterns that emerge, which corroborate observations from our
experiments with ABM. In particular, there is a general consensus about the best-performing
mechanisms. The MSE and $\text{MSE}_p$ baselines are among the best-performing mechanisms at nearly
every point. The best-performing peer prediction mechanisms from the simulated experiments,
$\Phi$-$\text{Div}_p$: KL and $\Phi$-$\text{Div}_p$: $H^2$, are also often among the best mechanisms, albeit less consistently.
Indeed, if we average the value of $\tau_b$ for each mechanism across all numbers of assignment blocks
and all semesters, we find that, the 5 best-performing mechanisms on average are $\text{MSE}_p$, MSE,
$\Phi$-$\text{Div}_p$: KL, OA, and $\Phi$-$\text{Div}_p$: $H^2$, in that order. These are exactly the same 5 mechanisms that
results from a similar analysis of the simulated data, albeit in a slightly different order. Notably,
these two parametric mechanisms perform especially well relative to the other mechanisms in the
first block of assignments, when the least amount of information is available to a mechanism. In
many courses, grades are assigned after each assignment and using only information from that

---

5 In general, this rule need not exclude students associated with few peer grades. In practice, however, it strikes a good
balance between being as inclusive as possible while excluding students associated with very few grades. In particular, it
seems to do better than using a threshold based on a percentage of the maximum number of peer grades for the given
semester.

6 In the experiments with real data, the differences between the values for $\Phi$-$\text{Div}_p$: KL, OA, and $\Phi$-$\text{Div}_p$: $H^2$ are not significant
enough to make reliable statements about their relative ordering.
Fig. 3. Quantifying Measurement Integrity with Real Data. For each semester, the average Kendall rank correlation coefficient ($\tau_B$) between students’ rewards and the average squared error of their reports from the true scores over 50 iterations of each mechanism.

particular assignment. In those cases, the amount of information available to a grading mechanism would be most similar to the amount of information available to the mechanisms in the first block of assignments in our experiments. This result mirrors the results from the experiments with ABM, in which the relative advantage of these two mechanisms over certain other peer prediction mechanisms (e.g., OA) tends to decrease as the number of assignments per semester increases (Figure 2).

5 RELATED WORK

Peer Prediction and Information Elicitation. For simplicity, we have so far used the generic term “peer prediction” primarily as a shorthand for a more specific setting—the elicitation of categorical signals without verification on multiple tasks, which was first explored (independently) by Dasgupta and Ghosh [2013] and Witkowski and Parkes [2013]. However, in the information elicitation literature, peer prediction mechanisms have been proposed in a variety of settings [Faltings et al., 2017b]. We choose to focus on one particular setting, because particularities of different settings necessitate different strategies for designing effective mechanisms. As a result, mechanisms for different settings will plausibly exhibit different empirical behaviors for properties like measurement integrity and robustness against strategic reporting. We leave the extension of our core ideas to other settings for future work.

The existing peer prediction literature, both in our setting and more broadly, is focused primarily on theoretical properties related to robustness against strategic behavior. However, there has been some work that explores incentive-compatibility from other perspectives. Gao et al. [2014] take an experimental approach and find evidence that agents are willing and able to exploit peer prediction mechanisms by coordinating on uninformative reports instead of truthful reports. Shnayer et al. [2016b] use replicator dynamics—a simulation-based approach that is quite different from our
approach with ABM and real data—to quantify desirable incentive properties of peer prediction mechanisms.

Some recent work has studied properties of peer prediction mechanisms beyond incentive properties. The result from this line of work most closely related to measurement integrity is by Kong [2020]. Kong establishes a theoretical “information evaluation” property for the DMI mechanism, which involves a theorem showing that (under certain assumptions) an agent’s expected payment under the DMI mechanism, given their signal, is proportional to a DMI-based measure of the quality of their report. However, this property still allows for the possibility that, due to noise, the relationship between rewards and qualities is not very strong at the population level. The DMI mechanism is not well-suited to the data-generating processes that we consider. However, under a data-generating process where data is more abundant, it would be relatively straightforward to adapt the components of the theoretical result to define a version of the information evaluation property as an instance of measurement integrity and explore the population-level relationship between rewards and qualities empirically.

Properties beyond incentives have also recently garnered attention in other information elicitation settings, many of which are distinct from the broader peer prediction paradigm because they assume that it is possible to access to ground truth information. In the setting of general crowdsourcing with limited access to ground truth, Goel and Faltings [2019] advance a novel notion of fairness—that the expected payment of each agent be directly proportional to the accuracy of their reports and independent from the strategy and accuracy of the other agents. The philosophical point embedded in this definition—that fair mechanisms must reward agents independently from the reports of other agents—would imply that any peer prediction mechanism is necessarily unfair. We do not accept this premise; our results demonstrate that, at least in some circumstances, certain peer prediction mechanisms have the ability to reliably reward agents fairly, even though they rely on the reports of other agents. In the forecasting setting, Li et al. [2022] and Neyman et al. [2021] both consider optimizing for properties related to incentivizing effort when selecting a proper scoring rule with which to score forecasts.

Another important work related to proper scoring rules is Liu et al. [2020], which considers the elicitation of forecasts (as opposed to the elicitation of categorical reports) without access to ground truth information. They propose a family of mechanisms, surrogate scoring rules (SSRs), which extend useful properties of proper scoring rules to the setting without verification. These properties include robustness against strategic behavior and “quantifying the value of information,” which is similar to Kong’s “information evaluation” property for the DMI mechanism, above. In addition to theoretical exploration of this latter property, they conduct experiments that quantify the extent to which the scores assigned by various mechanisms—including SSRs and certain peer prediction mechanisms from our setting (adapted to elicit forecasts instead of categorical reports)—correlate empirically with various metrics of forecast quality in each of a single draw from a variety of data-generating processes. Their experiments do not substantially differentiate the (adapted) peer prediction mechanisms they consider, nor do they suggest strategies by which to improve their performance. Lastly, they do not consider the interaction between the properties of robustness against strategic behavior and quantifying the value of information in practice.

Peer Prediction and Peer Assessment. There has also been work that explores the specific application of peer prediction to peer assessment. Shnayder and Parkes [2016] empirically analyze a limited set of peer prediction mechanisms using real MOOC data. Importantly, they show that some of the underlying assumptions made about data-generating processes in the theoretical literature—the self-dominating and self-predicting assumptions (see full version)—are likely to be violated in real peer assessment settings, especially when the report space is large. They also discuss the
importance of considering factors beyond expected rewards (e.g., the variance of rewards under a mechanism) in settings where fairness concerns are salient.

Radanovic et al. [2016] also apply peer prediction to peer assessment. In particular, they propose a particular mechanism, the Peer Truth Serum for Crowdsourcing (PTSC) mechanism⁷, and conduct an experiment where that mechanism and some baselines are used to reward peer graders with extra credit when grading a set of quizzes in a course on artificial intelligence at EPFL. They find that students who are rewarded using the PTSC mechanism grade more accurately than those rewarded using the baselines. However, they do not consider the relationship between the rewards and some measure of grading accuracy, nor whether the improvement in grading accuracy was the result of decreased strategic behavior or increased effort in grading.

More broadly, peer assessment is often touted as a natural application for the peer prediction paradigm. However, the typical approach in the literature has been to design mechanisms that are as generic as possible. The resulting mechanisms can be ill-suited to the challenges of a specific application, like those we identify for peer assessment in Section 1.3.1. Indeed, existing mechanisms often rely on collecting lots of data and compensating agents with rewards that exhibit high variance. Both these characteristics help explain why out-of-the-box peer prediction mechanisms tend to have low measurement integrity in our experiments.

On the other hand, certain works have been skeptical of the application of peer grading to peer assessment. Gao et al. [2016] and Zarkoob et al. [2020] explore how limited spot-checking (in the form of “ground-truth” grading by teaching assistants on certain assignments) can (theoretically) incentivize truthful reporting in grading in a simple model. Surprisingly, Gao et al. [2016] find that, compared with spot checking alone, supplementing spot-checking with peer prediction increases the number of spot-checks required to obtain the desired theoretical incentive properties. However, they do not consider how the objective of minimizing the number of spot checks may be in tension with rewarding students fairly. In more practical explorations of the utility of spot-checking, Wright et al. [2015] and Zarkoob et al. [2021] propose and refine, respectively, a peer grading system that is centered around the deployment of teaching assistants to improve the quality of feedback. We leave the application of our ideas to mechanisms that incorporate spot checking to future work.

**Modeling Peer Assessment.** The construction of our agent-based model is most influenced by the analysis of MOOC data on the platform Coursera conducted by Piech et al. [2013]. Piech et al. (and subsequent work, e.g., Zarkoob et al. [2022]) propose a sequence of increasingly complex parametric statistical models of peer assessment. Piech et al. show that estimating the parameters of each of their models is useful for estimating the true grades of student submissions that have been evaluated by peers. Grade estimates computed using their models are found to outperform grade estimates computed by the algorithm used by Coursera at the time. The inclusion of grader biases in each of their models is found to be the most significant single factor underlying this result.

Our model (Section 2.1) is not one proposed by Piech et al. (or their successors), but it is structurally similar to their model PG¹, which strikes a good balance between simplicity and performance in their analysis. The decision to propose a new model stems from a few important points: (1) Their models are continuous. In practice, though, essentially all assignment scores, rubrics, etc. are discrete. Thus, the “true grade” of a submission, and each grader’s signal, in a peer assessment model should be discrete. (2) Nearly all peer prediction mechanisms require a discrete report space. (3) It allows us to use model PG¹ in the implementation of our parametric peer prediction mechanisms (Section 3.3) without giving the mechanisms unrealistically accurate information about the underlying process by which true grades and reports are generated.

---

⁷We note that the PTSC mechanism is essentially a prototype of the PTS mechanism we consider; the latter is more suited to our setting.
6 DISCUSSION

We introduced measurement integrity as a novel property to consider in the design and analysis of peer prediction mechanisms. Alongside the more well-studied property of robustness against strategic reporting (see full version), measurement integrity plays an important role in understanding their practical performance. We focused on quantifying these properties empirically, using computational experiments with both an ABM and with real data. As a result, we were able to meaningfully differentiate mechanisms from the peer prediction literature in a way that has not been possible with theoretical analysis alone. Ultimately, we identified an apparent trade-off between our two dimensions of analysis (Figure 1) and found that parametric peer prediction mechanisms were best able to balance the two properties that characterize those dimensions.

Our unambiguous results suggest that our methodology—performing computational experiments to quantify mechanisms’ empirical properties—is useful for investigating desiderata, including, but not necessarily limited to, measurement integrity and robustness against strategic reporting. In particular, that our approach facilitates direct comparisons between mechanisms and uncovers consequences of implementation choices that are often abstracted away in theoretical analysis may be useful for selecting a peer prediction mechanism to deploy in a particular application.

Our results are driven by estimates of empirical quantities that are specific to the particular peer assessment data-generating processes that we consider. However, just as we find that they accord across these related data-generating processes, we expect many of those results to be relevant in other similar contexts. Our ABM is structurally similar to a model from the peer assessment literature that was validated using a large peer grading dataset from Massive Open Online Courses (MOOCs) by Piech et al. [2013]. As a result, it is reasonable to expect that the qualitative results of our experiments with that ABM should, at the very least, extend to that important class of peer assessment settings. The corroboration of these results by similar experiments with real peer grading data (not from a MOOC) suggests even broader applicability. Additionally, the relative scarcity of data for any particular grading task or agent in the peer assessment setting (Section 1.3.1) appears to be a significant driver of many of our results. This suggests that much of the intuition that we gain from our case study in peer assessment is likely to generalize to other settings where data is similarly sparse. Future theoretical work should explore the design of mechanisms—including parametric mechanisms, which seem well-suited for this purpose—that are less reliant on an abundance of data than the current mechanisms from the peer prediction literature.

We also expect the trade-off between measurement integrity and robustness against strategic behavior to remain in other settings. In the full version of the paper, we conduct experiments with mechanisms not yet studied in the peer prediction literature. We find the trade-off is persistent—no novel mechanisms significantly extend the Pareto frontier established by the mechanisms from the current peer prediction literature. On the other hand, particular challenges of peer assessment, like the scarcity of data, may not be a concern in some settings of interest. Thus, a mechanism’s performance in our experiments will not necessarily predict its performance universally.

ACKNOWLEDGMENTS

The authors would like to thank Professor Jason Hartline at Northwestern University for sharing peer grading data with us and Michalis Mamakos for helping us use it effectively. Thanks also to Kevin Leyton-Brown, Fang-Yi Yu, and Yichi Zhang for helpful discussions in developing the ideas explored in this work.

This work is supported by the National Science Foundation under award #2007256.
REFERENCES

Rémi Bachelet, Drissa Zongo, and Aline Bourelle. 2015. Does peer grading work? How to implement and improve it? Comparing instructor and peer assessment in MOOC GdP. In European MOOCs Stakeholders Summit 2015 (Proceedings of the Research Track). Mons, Belgium. https://halshs.archives-ouvertes.fr/halshs-01146710

Anujit Chakraborty, Jatin Jindal, and Swaprava Nath. 2020. Incentivizing Effort and Precision in Peer Grading. (2020). arXiv:1807.11657v6 [cs.GT] Working Paper.

Anirban Dasgupta and Arpita Ghosh. 2013. Crowdsourced Judgement Elicitation with Endogenous Proficiency. In Proceedings of the 22nd International Conference on World Wide Web (Rio de Janeiro, Brazil) (WWW ’13). Association for Computing Machinery, New York, NY, USA, 319–330. https://doi.org/10.1145/2488388.2488417 arXiv:1303.0799

Luca de Alfaro and Michael Shavlovsky. 2014. CrowdGrader: A Tool for Crowdsourcing the Evaluation of Homework Assignments. In Proceedings of the 45th ACM Technical Symposium on Computer Science Education (Atlanta, Georgia, USA) (SIGCSE ’14). Association for Computing Machinery, New York, NY, USA, 415–420. https://doi.org/10.1145/2538862.2538900 arXiv:1308.5273

Mikhail Drugov and Dmitry Ryvkin. 2020. How noise affects effort in tournaments. Journal of Economic Theory 188 (2020), 105065. https://doi.org/10.1016/j.jet.2020.105065

Boi Faltings, Radu Jurca, and Goran Radanovic. 2017a. Peer Truth Serum: Incentives for Crowdsourcing Measurements and Opinions. arXiv:1704.05269

Boi Faltings, Goran Radanovic, and Ronald Brachman. 2017b. Game Theory for Data Science: Eliciting Truthful Information. Morgan & Claypool Publishers.

Xi Alice Gao, Andrew Mao, Yiling Chen, and Ryan Prescott Adams. 2014. Trick or Treat: Putting Peer Prediction to the Test. In 15th ACM Conference on Economics and Computation (EC 2014). 507–524. https://doi.org/10.1145/2600057.2600265

Xi Alice Gao, James R. Wright, and Kevin Leyton-Brown. 2016. Incentivizing Evaluation via Limited Access to Ground Truth: Peer-Prediction Makes Things Worse. Workshop on Algorithmic Game Theory and Data Science at the 17th ACM Conference on Electronic Commerce. arXiv:1606.07042

Naman Goel and Boi Faltings. 2019. Deep Bayesian Trust: A Dominant and Fair Incentive Mechanism for Crowd. Proceedings of the AAAI Conference on Artificial Intelligence 33, 01 (Jul. 2019), 1996–2003. https://doi.org/10.1609/aaai.v33i01.33011996

Yuqing Kong. 2020. Dominantly Truthful Multi-task Peer Prediction with a Constant Number of Tasks. In Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms (SODA20). https://doi.org/10.1137/1.9781611975994.147 arXiv:1911.00272

Yuqing Kong and Grant Schoenebeck. 2019. An Information Theoretic Framework For Designing Information Elicitation Mechanisms That Reward Truth-telling. ACM Transactions on Economics and Computation 7, 1 (Jan. 2019). https://doi.org/10.1145/3296670 arXiv:1605.01021

Yingkai Li, Jason D. Hartline, Liren Shan, and Yifan Wu. 2022. Optimization of Scoring Rules. In Proceedings of the 23rd ACM Conference on Economics and Computation (Boulder, CO, USA) (EC ’22). Association for Computing Machinery, New York, NY, USA, 988–997. https://doi.org/10.1145/3490486.3538338 arXiv:2007.02905

Yang Liu, Juntao Wang, and Yiling Chen. 2020. Surrogate Scoring Rules. In Proceedings of the 21st ACM Conference on Economics and Computation (Virtual Event, Hungary) (EC ’20). Association for Computing Machinery, New York, NY, USA, 853–871. https://doi.org/10.1145/3391403.3399488 arXiv:1802.09158

Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting Informative Feedback: The Peer-Prediction Method. Management Science 51, 9 (2005), 1359–1373. https://doi.org/10.1287/mnsc.1050.0379

Eric Neyman, Georgy Noarov, and S. Matthew Weinberg. 2021. Binary Scoring Rules That Incentivize Precision. In Proceedings of the 22nd ACM Conference on Economics and Computation (Budapest, Hungary) (EC ’21). Association for Computing Machinery, New York, NY, USA, 718–733. https://doi.org/10.1145/3465456.3467639 arXiv:2002.10669

Chris Piech, Jonathan Huang, Zhenghao Chen, Chuang Do, Andrew Ng, and Daphne Koller. 2013. Tuned Models of Peer Assessment in MOOCs. In 6th International Conference on Educational Data Mining, Memphis, TN. arXiv:1307.2579

Goran Radanovic, Boi Faltings, and Radu Jurca. 2016. Incentives for Effort in Crowdsourcing Using the Peer Truth Serum. ACM Trans. Intell. Syst. Technol. 7, 4, Article 48 (mar 2016), 28 pages. https://doi.org/10.1145/2856102

Thomas L. Saaty. 2004. Scales from measurement–not measurement from scales. In Proceedings of the 17th International Conference on Multiple Criteria Decision Making, Whistler, BC Canada. 6–11.

Grant Schoenebeck and Fang-Yi Yu. 2021. Learning and Strongly Truthful Multi-Task Peer Prediction: A Variational Approach. In 12th Innovations in Theoretical Computer Science Conference (ITCS 2021). arXiv:2009.14730

Victor Shinayder, Arpit Agarwal, Rafael Frongillo, and David C. Parkes. 2016a. Informed Truthfulness in Multi-Task Peer Prediction. In 17th ACM Conference on Economics and Computation (EC 2016). 179–196. https://doi.org/10.1145/2940716.2940790 arXiv:1603.03151

Victor Shinayder, Rafael M. Frongillo, and David C. Parkes. 2016b. Measuring Performance of Peer Prediction Mechanisms Using Replicator Dynamics. In 25th International Joint Conference on Artificial Intelligence (IJCAI 2016). New York, NY. http://nrs.harvard.edu/urn-3:HUL.InstRepos:32220916
Victor Shnayder and David C. Parkes. 2016. Practical Peer Prediction for Peer Assessment. In Fourth AAAI Conference on Human Computation and Crowdsourcing, Austin, TX. http://nrs.harvard.edu/urn-3:HUL.InstRepos:34732142

S. S. Stevens. 1946. On the Theory of Scales of Measurement. Science 103, 2684 (1946), 677–680. http://www.jstor.org/stable/1671815

Jens Witkowski and David C. Parkes. 2013. Learning the Prior in Minimal Peer Prediction. In 3rd Workshop on Social Computing and User Generated Content (SC 2013). Philadelphia, PA, 39:1–39:12. http://nrs.harvard.edu/urn-3:HUL.InstRepos:34222829

James R. Wright, Chris Thornton, and Kevin Leyton-Brown. 2015. Mechanical TA: Partially Automated High-Stakes Peer Grading. In 46th ACM Technical Symposium on Computer Science Education. 96–101. https://doi.org/10.1145/2676723.2677278

Zheng Yuan and Doug Downey. 2020. Practical Methods for Semi-Automated Peer Grading in a Classroom Setting. Association for Computing Machinery, New York, NY, USA, 363–367. https://doi.org/10.1145/3340631.3394878

Hedayat Zarkoob, Farzad Abdolhosseini, and Kevin Leyton-Brown. 2021. Mechanical TA 2: A System for Peer Grading with TA Support. arXiv:2101.10078 [cs.CY]

Hedayat Zarkoob, Greg d’Eon, Lena Podina, and Kevin Leyton-Brown. 2022. Better Peer Grading through Bayesian Inference. arXiv:2209.01242 [cs.AI]

Hedayat Zarkoob, Hu Fu, and Kevin Leyton-Brown. 2020. Report-Sensitive Spot-Checking in Peer-Grading Systems. In Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (Auckland, New Zealand) (AAMAS ’20). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1593–1601. arXiv:1906.05884