Spot Check Equivalence: an Interpretable Metric for Information Elicitation Mechanisms

Shengwei Xu*
shengwei@umich.edu
University of Michigan
Ann Arbor, USA

Yichi Zhang
yichiz@umich.edu
University of Michigan
Ann Arbor, USA

Paul Resnick
presnick@umich.edu
University of Michigan
Ann Arbor, USA

Grant Schoenebeck
schoeneb@umich.edu
University of Michigan
Ann Arbor, USA

ABSTRACT

Because high-quality data is like oxygen for AI systems, effectively eliciting information from crowdsourcing workers has become a first-order problem for developing high-performance machine learning algorithms. Two prevalent paradigms, spot-checking and peer prediction, enable the design of mechanisms to evaluate and incentivize high-quality data from human labelers. So far, at least three metrics have been proposed to compare the performances of these techniques [2, 6, 31]. However, different metrics lead to divergent and even contradictory results in various contexts. In this paper, we harmonize these divergent stories, showing that two of these metrics are actually the same within certain contexts and explain the divergence of the third. Moreover, we unify these different contexts by introducing *Spot Check Equivalence*, which offers an interpretable metric for the effectiveness of a peer prediction mechanism. Finally, we present two approaches to compute spot check equivalence in various contexts, where simulation results verify the effectiveness of our proposed metric.

CCS CONCEPTS

• **Theory of computation** → *Algorithmic game theory and mechanism design.*

KEYWORDS

Algorithmic Game Theory, Information Elicitation, Incentive for Effort, Peer Prediction

ACM Reference Format:
Shengwei Xu, Yichi Zhang, Paul Resnick, and Grant Schoenebeck. 2024. Spot Check Equivalence: an Interpretable Metric for Information Elicitation Mechanisms. In *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3589334.3645679

*Corresponding author
†This work is supported by United States National Science Foundation award number 2313137 and 2007256

WWW '24, May 13–17, 2024, Singapore, Singapore.

ACM ISBN 979-8-4007-0171-9/24/05…$15.00
https://doi.org/10.1145/3589334.3645679

1 INTRODUCTION

Eliciting precise and valuable information from individuals is becoming paramount, especially with the rising demands for data labeling in the realms of AI and machine learning. Recent advancements, such as Large Language Models (LLMs), have proven the value of high-quality human-labeled data. For example, Meta is estimated to have invested upward of 25 million dollars in collecting preference data from human labelers to align Llama 2 with human preferences [17]. This raises a pressing question: How can human agents be incentivized to provide high-quality information. E.g., without the proper incentives, human labelers for LLM alignment may not exert effort to distinguish between truthful LLM responses and merely authoritative-sounding responses (i.e. hallucinations), even when truthfulness is important for the task at hand.

Research from Amazon reveals that monetary compensation is the principal motivator for Amazon Mechanical Turk workers [3], and indeed the primary solution is to monetarily reward agents in exchange for effortful and truthful labeling. Two distinct compensation strategies, *spot-checking* [10] and *peer prediction* [19], each rate the quality of user feedback with a score. This score can then be transformed into a payment for an agent.

Spot-checking mechanisms reward agents by comparing their reports with the ground truth on a small fraction of gold standard questions. When the ground truth information is expensive or even infeasible to obtain, peer prediction mechanisms are proposed, which reward an agent based on the correlation between her reports and the reports of other agents.

To understand and compare the performance of mechanisms developed from these paradigms, there is a need for standard metrics similar to accuracy, recall, and F1 score used in supervised learning. Notice that all these metrics range from 0 to 1 where 1 is good/perfect and 0 is bad. While several studies have proposed methods for comparing these mechanisms [2, 6, 31], there remains a conspicuous gap for both a unified understanding of how these metrics relate, and, if possible, a unified interpretable metric.

To this end, we introduce the concept of *Spot Check Equivalence* (SCE), which uses a spot-checking mechanism as a benchmark to quantify the *motivational proficiency* of an arbitrary incentive mechanism. As introduced in Zhang and Schoenebeck [31],

the motivational proficiency is the minimum cost of budget to induce a desired effort level in a symmetric equilibrium. Then, a SCE of 1 will indicate that a mechanism does as well as a certain spot-checking mechanism does when the spot-checking mechanism has access to the ground truth of every task. A SCE of 0 will indicate that a mechanism does as well as this same spot-checking mechanism when it has no access to the ground truth of any task (essentially, it is paying agents randomly). Note that accessing the ground truth might be costly, e.g., hiring an expert to get the ground truth might be much more expensive than hiring several non-expert crowd workers. Thus, SCE can quantify the considerable cost savings that might be achieved by employing a peer prediction mechanism over a straightforward spot-checking mechanism.

By sufficiently harshly punishing the agents for the checked low-quality reports, spot-checking mechanisms can effectively motivate effort, even when only a small fraction of the tasks are checked. However, in most real applications, the payoff should be non-negative which precludes this approach.

Gao et al. [6, 7] study a peer grading setting where agents are modeled as having a binary choice for the effort to exert: low versus high. In their setting, the goal is to minimize the fraction of random questions that must be spot-checked to incentivize agents to exert high effort (make choosing high effort a Nash equilibrium). They find that, in their model, combining spot-checking with peer prediction does not help reduce the spot-checking ratio required to achieve the desired incentive properties, i.e. peer prediction makes things worse. However, their results rest on several assumptions, which we will discuss later in Section 6.

Burrell and Schoenebeck [2] propose a metric called *Measurement Integrity* to quantify the ex-post fairness of a peer prediction mechanism. Mechanisms with high Measurement Integrity can produce payments that are strongly correlated with the quality of the agents' reports. Their motivation and definitions look beyond incentives to fairness. They do not study spot-checking mechanisms, but it is clear that the more agents are spot-checked the more accurately their scores will reflect their true quality. For example, with ground truth for all the tasks, an agent's score could exactly reflect the true quality of her responses. Moreover, they model continuous effort but do not establish a clear link between Measurement Integrity and the ability to incentivize effort.

Zhang and Schoenebeck [31] study incentivizing effort in a crowdsourcing setting where agents can choose their effort from a continuum. Their goal is to maximize the motivational proficiency by rescaling the scores output by the incentive mechanism into practical payments. They suggest a tournament-based payment scheme: first rank the agents based on their scores output by an incentive mechanism, and then pay a predetermined reward for each ranking. Under the tournament setting, they further propose a sufficient statistic of a mechanism's motivational proficiency, called the *Sensitivity*. Intuitively, the Sensitivity measures how responsive a score is to changes in an agent's effort. For example, a spot-checking mechanism that checks a larger fraction of tasks is more sensitive to changes in an agent's effort (intuitively, it has more chances to detect a change), and thus has a larger motivational proficiency. However, there is a lack of discussions on how to estimate Sensitivity in practice.

An apparent contradiction arises. Zhang and Schoenebeck [31] empirically show that when agents exert a reasonably high effort, peer prediction mechanisms have Sensitivity competitive with spot-checking mechanisms that randomly check 20% of the tasks. However, the aforementioned implication of Gao et al. [6, 7] would seem to predict that the spot-checking mechanisms are always superior to peer-prediction mechanisms.

Our contributions. First, we propose Spot Check Equivalence which uses the equivalent spot-checking ratio as an interpretable way to measure an information elicitation mechanism's performance under specified information elicitation contexts. We study the Spot Check Equivalence based on Measurement Integrity and Sensitivity, and demonstrate its effectiveness as a metric for motivational proficiency both theoretically and empirically.

Second, we unify Measurement Integrity (the metric for ex-post fairness) and Sensitivity (the metric that serves as a proxy for motivational proficiency). In particular, we prove that Spot Check Equivalence based on Measurement Integrity and Sensitivity are sometimes exactly the same. We also show why these results differ so much from Gao et al. [6, 7], and thus refute, or at least qualify, the titular statement that "Peer-prediction makes things worse."

Third, we present two approaches to compute Spot Check Equivalence, which are suitable for settings with and without ground truth data, respectively. Our method enables the comparison of the motivational proficiency of different mechanisms across various information elicitation contexts. Furthermore, our simulation results show that both approaches result in similar estimations of SCE, which implies the robustness of our methods.

2 MODEL

In this section, we will give a formal definition of the information elicitation context (Figure 1), and then formally define the Spot Check Equivalence.

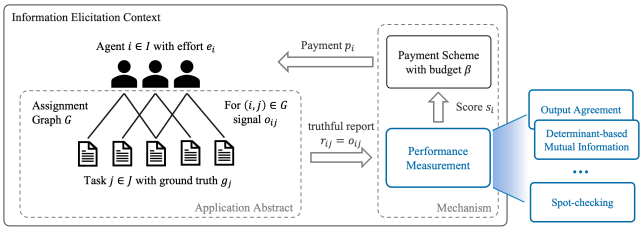


Figure 1: Information Elicitation Context

2.1 Information Elicitation Context

Formally, as shown in Figure 1, an information elicitation context (IEC) is defined as a tuple:

Information Elicitation Context $(IEC) := (Agent, App, Mech)$

where $Agent = (I, c, \mathbf{e})$ represents the agents and their properties, $App = (J, \mathcal{GT}, \omega, \Sigma, D)$ represents an information elicitation application abstraction, and $Mech = (M, P)$ represents a mechanism. We assume that the information elicitation context is common knowledge for all the agents.

Agent. In $Agent = (I, c, \mathbf{e})$: I is the set of agents; $\mathbf{e} = [e_i]_{i\in I} \in [0,1]^{|I|}$ represents all agents' effort levels. the cost function $c : [0,1] \to \mathbb{R}^+ \cup \{0\}$ maps an effort level to a non-negative, increasing, and convex cost. Notice agents are homogeneous and share the same cost function.

Application Abstraction. $App = (J, \mathcal{GT}, \omega, \Sigma, D)$ comprises the task set J, the ground truth space $\mathcal{GT}$, the prior of the ground truth $\omega = \Delta_{\mathcal{GT}}$, the signal space Σ, and the data-generating process $D = (D_{assign}, D_{signal})$.

D_{assign} describes how the tasks are assigned to the agents.

$$D_{assign} : \Delta_{\mathcal{G}}$$

where $\mathcal{G}$ represents the space over G, and $G = (I \cup J, E_G)$ represents a bipartite graph between I and J, indicating how the tasks are assigned to the agents.

D_{signal} describes how the signals are generated: the distribution of agent i's signal on task j conditioned on the effort $e_i \in [0,1]$ and task j's ground truth $g_j \in \mathcal{GT}$, given the edge $(i,j) \in E_G$:

$$D_{signal} : [0,1] \times \mathcal{GT} \to \Delta\Sigma$$

Agents' Report. We assume agents truthfully report the signals they obtain from the application abstraction conditioned on their effort levels. Further discussion will be provided in Appendix B. And we denote the agent i's report on task j as $r_{ij} \in \Sigma$.

Application Instance. With the specified $Agent = (I, c, \mathbf{e})$ and $App = (J, \mathcal{GT}, \omega, \Sigma, D)$, we can generate an instance representing a realized information elicitation application:

- For the given I, J, we sample an assignment graph G according to D_{assign}.
- For each task $j \in J$, we independently sample its ground truth g_j from the prior ω.
- For each pair $(i,j) \in E_G$, we independently sample agent i's signal on task j from the distribution $D_{signal}(e_i, g_j)$, denoted as $o_{ij} \in \Sigma$.
- For each pair $(i,j) \in E_G$, as we assumed, the agent i's report $r_{ij} = o_{ij}$.

The mechanism takes the application instance as input.

Performance Measurement. The performance measurement M is a component of the mechanism $Mech = (M, P)$. It maps the agents' reports to their performance scores. Formally,

$$\text{(Peer Prediction) } M : \Sigma^{|E_G|} \to_{random} \mathbb{R}^{|I|}$$

$$\text{(Spot-checking) } M : \Sigma^{|E_G|} \times \mathcal{GT}^{|J_{\text{checked}}|} \to_{random} \mathbb{R}^{|I|}$$

Note that the spot-checking performance measurement can access the ground truth of the checked tasks $J_{\text{checked}} \subseteq J$.

We use $s_i \in \mathbb{R}$ to denote agent i's score, and $\mathbf{s} = [s_i]_{i\in I}$ to denote the vector of all agents' scores.

Payment Scheme. The payment scheme P is the other component of the mechanism. It maps the agents' performance scores to the payoffs, which are directly related to their utilities. Formally,

$$\text{(Payment Scheme) } P : \mathbb{R}^{|I|} \to \left(\mathbb{R}^* \cup \{0\}\right)^{|I|}$$

We use $p_i \in \mathbb{R}^* \cup \{0\}$ to denote payoff of agent i, and $\beta = \sum_{i\in I} p_i$ to denote the total payment among all the agents. And we denote the vector of all the agents' payoffs as $\mathbf{p} = [p_i]_{i\in I}$.

For intuition, we give two examples of payment schemes:

Definition 2.1 (Linear Payment Scheme). *A linear payment scheme pays a payoff $p_i = a \cdot s_i + b$ to each agent i, where a, b are the constant parameters.*

Definition 2.2 (Tournament Payment Scheme). *A tournament payment scheme ranks the agents according to their scores and pays the i-th ranked agent $\hat{p_i}$, where $\hat{p_1}, \hat{p_2}, ..., \hat{p_{|I|}}$ are constant parameters that monotonically decreasing, i.e., $\hat{p_i} \geq \hat{p_{i'}}$ when $i \leq i'$. Without loss of generality, we assume $s_1 \geq s_2 \geq ... \geq s_{|I|}$, and thus $p_i = \hat{p_i}$.*

Report quality. Given an instance, we can define the quality of an agent's report. The quality function Q for one report is a deterministic loss function:

$$\text{(Quality Function) } Q : \Sigma \times \mathcal{GT} \to \mathbb{R}$$

Agent i's overall report quality q_i is defined as the average of her reports' qualities, i.e. $q_i = \sum_{j|(i,j)\in G} Q(r_{ij}, g_j)$. We denote the vector of all the agents' qualities as $\mathbf{q} = [q_i]_{i\in I}$.

Equilibrium. We assume that the agents choose their effort level according to the following equilibrium.

Definition 2.3 (Symmetric local equilibrium). *Given an IEC where all the agents exert effort $e_i = \xi$ and $\mathbb{E}[\sum_{i\in I} p_i] \geq |I| \cdot c(\xi)$ (Individual Rationality is satisfied), we say it is a symmetric local equilibrium if the derivative of every agent's utility is 0, i.e.*

$$\frac{\partial}{\partial e_i} u(e_i, e_{-i} = \xi)|_{e_i=\xi} = 0$$

Note that, at this equilibrium, $\frac{\partial}{\partial e_i} u(e_i, e_{-i} = \xi)|_{e_i=\xi} = 0$ is a necessary condition for ξ being a Nash equilibrium. Zhang and Schoenebeck [31] show empirical evidence that a local equilibrium is very likely to be a Nash equilibrium in the model we mainly discuss in our paper. In the rest of our paper, our discussion will focus on this equilibrium.

2.2 Motivational Proficiency

The motivational proficiency of a component (a mechanism, a performance measurement, or a payment scheme) within an information elicitation context *IEC* represents its ability to incentivize effort. To quantify it, we fix all the other components of the *IEC* and quantify the expected total payment for eliciting a fixed effort level at the equilibrium (Definition 2.3), lower expected total payment implies higher motivational proficiency.

Definition 2.4 (Motivational proficiency). *We define the motivational proficiency of a component (Mech, M, or P) within an information elicitation context IEC where all the agents exert effort level ξ as the negative expected total payment needed to realize the symmetric local equilibrium (Definition 2.3) at effort level ξ when substituting the component into the information elicitation context.*

As we discussed in the introduction, Zhang and Schoenebeck [31] show that tournament payment schemes have higher motivational proficiency compared to linear payment schemes in certain settings where limited liability is needed. Therefore, in the following discussion, we will focus on the motivational proficiency of performance measurements in information elicitation contexts with

a tournament payment scheme. In Section 5, we quantify the total payment in an information elicitation context with tournament payment. A simple example (Example A) in Appendix A illustrates the intuition behind motivational proficiency and many other concepts of this paper.

2.3 Measure of Performance Measurements

In addition to motivational proficiency, there are other measures of performance measurements, as we discussed in the introduction, including Sensitivity [31] and Measurement Integrity [2]. We now propose the general definition.

Definition 2.5 (Measure of Performance Measurement). *A measure f of performance measurement M within information elicitation context IEC maps the IEC with M to a real number, denoted as $f(IEC \leftarrow M) \in \mathbb{R}$, where the leftarrow means we apply M in the information elicitation context IEC.*

We now show the two examples of measure f, Sensitivity [31] and Measurement Integrity [2].

Sensitivity. Zhang and Schoenebeck [31] propose the Sensitivity as a proxy of the motivational proficiency of a performance measurement. They show that the motivational proficiency highly depends on the Sensitivity (Definition 2.6, Lemma 2.7), which measures how an agent's performance score changes when she deviates from the equilibrium effort level ξ. When all other agents exert effort ξ, we denote the mean of agent i's score as $\mu_s(e_i)$, and the standard deviation as $\sigma_s(e_i)$.

Definition 2.6 (Sensitivity [31]). *The Sensitivity of a performance measurement within an information elicitation context IEC at equilibrium effort level ξ is defined as*

$$\text{Sensitivity}(IEC \leftarrow M) = \delta(\xi) = \frac{\frac{\partial}{\partial e_i}\mu_s(e_i)|_{e_i=\xi}}{\sigma_s(\xi)}$$

Lemma 2.7 (Proposition 4.8 in Zhang and Schoenebeck [31]). *If the agent i's score s_i follows a normal distribution $N(\mu_s(e_i), \sigma_s(e_i)^2)$, the expected total payment to elicit effort ξ will (weakly) decrease in the Sensitivity $\delta(\xi)$ in a specific information elicitation context with any tournament payment scheme.*

Measurement Integrity. To measure the ex-post fairness of a performance measurement, Burrell and Schoenebeck [2] propose the Measurement Integrity, which is defined as the expected correlation between the quality of the agents' reports and their performance scores.

Definition 2.8 (Measurement Integrity [2]). *Formally, the Measurement Integrity of a performance measurement M with respect to a quality function Q and a correlation function* corr*, within an information elicitation context IEC is*

$$\underset{Q,\text{corr}}{\text{MI}}(IEC \leftarrow M) = \mathbb{E}_{IEC}[\text{corr}(\mathbf{s}, \mathbf{q})]$$

2.4 Using Spot-checking as Reference: Spot Check Equivalence

Even though we have metrics which can compare two performance measurements, there is still a need for a metric for a performance measurement whose value is per se. Our idea is to take any metric of performance measurements, f, and convert it to an interpretable metric as follows: instead of using the actual value of f applied to some performance measurement, instead use the checking ratio of the spot-checking performance measurement who, when also evaluated by f, yields a value equivalent to that of the performance measurement.

First, we adopt the definition of spot-checking performance measurement from Gao et al. [6], and assume it can access the ground truth, thus, it is an idealized spot-checking. Given that we exclusively focus on this particular spot-checking, we might omit the term 'idealized' in subsequent discussions for brevity.

Definition 2.9 (Spot-checking performance measurement (idealized)). *An (idealized) spot-checking performance measurement is denoted as a tuple $SC := (X, S, C)$, which checks X percent of all the tasks u.a.r.; then scores the agent i with $S(r_{ij}, g_j)$ for each checked task j, and score $C \in \mathbb{R}$ for each of the unchecked tasks.*

Intuitively, a higher checking ratio leads to less noise, so high effort is easier to notice, which is beneficial for both motivational proficiency and ex-post fairness. Thus, we can use the equivalent spot-checking ratio as a metric for both motivational proficiency and ex-post fairness of a performance measurement.

Formally, we can define the Spot Check Equivalence as follows.

Definition 2.10 (Spot Check Equivalence). *For a performance measurement M within an information elicitation context $IEC = (Agent, App, Mech)$, at the symmetric local equilibrium $\mathbf{e} = \xi$, given a measure of performance measurements f (Definition **??**), we define the Spot Check Equivalence SCE of M, with respect to a spot-checking mechanism SC as*

$$SCE(IEC \leftarrow M) = \sup_X\{f(IEC \leftarrow M) \geq f(IEC \leftarrow SC(X, S, C))\}$$

In our following discussion, to make the spot-checking mechanism $SC(X, S, C)$ non-trivial[1], we use a quality function to score the agents for checked tasks, i.e. $S = Q$. And since we will apply a payment scheme after the performance measurement, the value of the constant score C for unchecked tasks does not matter, thus, we set $C = 0$. We then use $SC(X)$ to denote $SC(X, S = Q, C = 0)$.

In the next section, we will theoretically prove the unification of Sensitivity and Measurement Integrity in certain settings, and consequently, we can show that the Spot Check Equivalence based on the Sensitivity or Measurement Integrity can be used as an interpretable metric for the motivation proficiency of an information elicitation performance measurement. We will show more empirical evidence in our agent-based model simulations (Section 5) when relaxing the theoretical assumptions.

3 UNIFICATION OF SENSITIVITY AND MEASUREMENT INTEGRITY

In this section, we formally prove that the Spot Check Equivalence based on Sensitivity or Measurement Integrity can be used as a proxy for Spot Check Equivalence based on motivational proficiency in certain settings. Given that computing these three measures has different requirements, the unification allows us to compute Spot Check Equivalence and consequently, measure the motivational proficiency in more scenarios.

[1] An example of a trivial spot-checking mechanism is $S(r_{ij}, g_j) \equiv 0$ and $C \equiv 0$.

We formally propose and prove the unification of Measurement Integrity and Sensitivity in our main theorem (Theorem 3.4). Recall that Zhang and Schoenebeck [31] have shown that, within certain settings, high Sensitivity leads to low expected total payment (Lemma 2.7) when applying a tournament payment scheme. Putting these together, we have that the Spot Check Equivalence based on motivational proficiency, Sensitivity, and Measurement Integrity are equal.

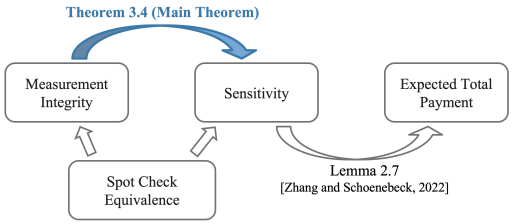


Figure 2: Theoretical Analysis Overview

Since the Sensitivity relies on the distribution of the performance score, we make the following assumptions about the distributions of the quantities in our model, similar to prior work [31].

Assumption 3.1 (The Gaussian assumption for the quality). *Given effort level e_i, the quality q_i follows a normal distribution $N(e_i, \sigma_q(e_i)^2)$. And we further assume that $\sigma'_q(e_i) << \sigma_q(e_i)$.*

Assumption 3.2 (The Gaussian assumption for the score). *Given the report quality q_i, the score s_i follows a normal distribution $N(\mu_{s|q}(q_i), \sigma_{s|q}(q_i)^2)$.*

Assumption 3.3 (The independent assumption for score). *When all agents have the same effort level $e_i = \xi$ and the number of agents $|I|$ goes to infinity, the agents' performance scores are independent.*

Theorem 3.4 (Main Theorem). *For a given performance measurement M within an information elicitation context IEC where every agent exert effort level ξ, when Assumption 3.1 3.2 and 3.3 are satisfied, there exists a linear bijection between the* $\mathrm{MI}_{Q,\mathrm{corr}}(IEC \leftarrow M)$ *and the Sensitivity $\delta(\xi)$, where* corr *is the sample Pearson correlation coefficient and the number of agents goes to infinity.*

We further discuss the reasonability of the assumptions and the proof of our main theorem in our arXiv version.

Note that our main theorem (Theorem 3.4) relies on Assumption 3.1, 3.2, and 3.3, thus, it is important to examine whether the unification of motivational proficiency, Sensitivity, and Measurement Integrity is still true with real scores calculated by various performance measurements. In Section 5, we will demonstrate some positive evidence from our agent-based model experiment, where Assumption 3.1, 3.2, and 3.3 are relaxed.

4 COMPUTING SPOT CHECK EQUIVALENCE

Given the unification of Sensitivity and Measurement Integrity, if we can compute the Sensitivity or Measurement Integrity of a performance measurement, we can get the Sensitivity-based or Measurement Integrity-based Spot Check Equivalence respectively.

If the measure $f(IEC \leftarrow SC(X))$ is monotonic, we can use a binary search algorithm (Algorithm 2 in Appendix C) to compute the Spot Check Equivalence.

4.1 Computation with Ground Truth

We first propose a workflow to compute the Spot Check Equivalence with the ground truth of the tasks. The Spot Check Equivalence is like accuracy, recall, and F1 score in machine learning, which can only be calculated on training or testing datasets rather than real applications. Similarly, it is reasonable to create datasets to evaluate the information elicitation performance measurements, get a good sense of their motivational proficiency, and then decide which performance measurement to apply in real applications.

With the ground truth of the tasks, we can calculate the quality of the reports, and then, use the correlation between the agents' scores and qualities to estimate the measurement integrity of the performance measurement M and $SC(X)$.

Intuitively, as more tasks are checked, the score s_i will be more correlated to the quality q_i. Our agent-based model experiment also shows that the Measurement Integrity monotonically increases with the spot-checking ratio (Section 5.2). Therefore, we can apply Algorithm 2 to estimate the Spot Check Equivalence based on Measurement Integrity.

4.2 Computation without Ground Truth

Considering the current lack of information elicitation dataset, we propose another method to estimate the Spot Check Equivalence without the ground truth.

Recall the definition of Sensitivity (Definition 2.6), both the derivative of the performance score $\frac{\partial}{\partial e_i}\mu_s(e_i)|_{e_i=\xi}$ and the standard deviation $\sigma_s(\xi)$ do not require access to the ground truth. The standard deviation $\sigma_s(\xi)$ can be estimated by the standard deviation of $\{s_i\}$ given that all the agents are homogeneous. However, in real data, to estimate $\frac{\partial}{\partial e_i}\mu_s(e_i)|_{e_i=\xi}$ is tricky because the score after deviating from ξ is not accessible.

We adopt the idea of bootstrap sampling: we randomly select an agent i, and if we know how the effort impacts the report, we can randomly manipulate her report as an intentional degradation of her effort by some ε. We then compute the mean difference between all selected agents' scores before and after the manipulation, denoted as $\Delta\mu$. Note that the mean difference $\Delta\mu$ can be regarded as an estimation of $\varepsilon \cdot \frac{\partial}{\partial e_i}\mu_s(e_i)|_{e_i=\xi}$, and consequently, $\Delta\mu/\sigma_s$ is proportional to the Sensitivity. For example, if decreasing the effort brings uniform noise, we get the Algorithm 1. In Section **??**, we present evidence demonstrating that Algorithm 1 works in our agent-based model simulation.

5 EFFECTIVENESS OF THE SPOT CHECK EQUIVALENCE: AGENT-BASED MODEL EXPERIMENT

In this section, we present the results of agent-based model (ABM) experiments to evaluate the effectiveness of the *Spot Check Equivalence* based on Measurement Integrity and Sensitivity in measuring the motivational proficiency, without the assumptions we made in our theoretical proof. We then further compare different peer

Algorithm 1: Estimate SCE without Ground Truth

Input: Information Elicitation Context *IEC*, Performance Measurement M, Iteration Times T
Output: Spot Check Equivalence *SCE*
Function *Estimate* $\Delta\mu(IEC \leftarrow M)$**:**
  $\Delta\mu = 0$
  for $t = 1$ **to** T **do**
    Choose $i \in I$ u.a.r.
    Compute score s_i with M
    for $(i, j) \in \mathcal{G}$ **do**
      if $random(0, 1) > \varepsilon$ **then**
        Choose $r_{ij} \in \mathcal{GT}$ u.a.r.
      end
    end
    Compute score s'_i with M
    $\Delta\mu = \Delta\mu + (s'_i - s_i)/T$
  end
  return $\Delta\mu$
$SCE = \texttt{BinarySearchSCE}(M, f = \Delta\mu/\sigma_s)$

prediction performance measurements with spot-checking performance measurements in different information elicitation contexts.

5.1 Model Setup

We first briefly introduce our agent-based model setup. The detailed setup is discussed in our arXiv version. According to the definition of the information elicitation context in Section 2, our agent-based model contains the following components:

Agents. We consider a population of $|I| = 50$ agents. Each agent i has an effort level $e_i = \xi \in [0, 1]$ with an associated cost function $c(e_i) = e_i^2$.

Data-generating Process (for Application Instance). We consider a context with $|J| = 500$ tasks. Each agent is assigned to 50 tasks, while each task is assigned to 5 agents.

We adopt a generalized Dawid-Skene model from previous work [31]. The ground truth $g_j \in \mathcal{GT}$ of task j is independently sampled from a discrete prior distribution ω learned from a dataset of a crowdsourcing task on Amazon Mechanical Turk [25, 31], where $\mathcal{GT}$ is a finite set including all possible ground truths.

The agent i will receive a signal $o_{ij} \in \Sigma$ on task j given her effort level and the task's ground truth. In our experiment, we assume that $\Sigma = \mathcal{GT}$. Then, we can define two $|\mathcal{GT}| \times |\Sigma|$ confusion matrices, Γ_{work} and Γ_{shirk}. The (row, col) entry of Γ_{work} and Γ_{shirk} represents the probability of getting the row-th signal conditioned on the col-th ground truth when the agent exerts effort level 1 and 0 respectively. When the agent i exert e_i effort, the confusion matrix is

$$\Gamma_i = e_i * \Gamma_{work} + (1 - e_i) * \Gamma_{shirk}$$

where the confusion matrix Γ_{work} is also learned from the above dataset [25, 31], and we set Γ_{shirk} as a matrix representing a uniform distribution.

Performance Measurement. We implement several peer prediction mechanisms, including Output Agreement (OA) [26–28], Peer Truth Serum (PTS) [5], Correlated Agreement (CA) [24], f-Mutual Information (f-MI) [15, 16], and Determinant-based Mutual Information (DMI)[2] [12], which yield different Spot Check Equivalences within the above information elicitation context.

Payment Scheme. We now introduce the tournament payment scheme we use in our simulation.

Borda-count payment scheme. A very intuitive way to pay an agent according to her ranking is to pay her how many agents she beats. When there is a draw, we split the payoff evenly. Formally, we have

$$p_i = C \cdot \#\text{beaten} = C \sum_{i' \in |I|, i' \neq i} \left(\mathbb{1}[s_i > s_{i'}] + \frac{1}{2} \mathbb{1}[s_i = s_{i'}] \right)$$

where C is a constant parameter and the total payment is $C \times \binom{|I|}{2}$.

To calculate the total payment[3] in the Borda-count scheme for a specific performance measurement M for the symmetric local equilibrium where every agent exerts $e_i = \xi$ effort. We let the derivative of the agent i's expected payoff equal to 0, which implies the parameter C should be

$$C = \frac{\frac{\partial}{\partial e_i} \mathbb{E}[\#\text{beaten} | e_i, e_{-i} = \xi]|_{e_i = \xi}}{\frac{\partial}{\partial e_i} c(e_i)|_{e_i = \xi}}$$

Note that to guarantee Individual Rationality (The agents' expected utility is non-negative), when the calculated optimal payment is less than the total cost of all the agents, we set the total payment as the total cost. We assume that if a payment scheme can incentivize effort level ξ using the optimal payment, it can also incentivize the same effort when the total payment is greater than the optimal payment.

5.2 Measurement-Integrity-based Spot Check Equivalence.

We examine whether the Spot Check Equivalence based on Measurement Integrity can indicate a performance measurement's motivational proficiency. Recall that the motivational proficiency of a performance measurement could be quantified by the amount of the expected total payment we need to elicit a fixed effort level in an information elicitation context. Thus, in the experiment examining the effectiveness, we mainly investigate the relationship between the Measurement Integrity and the expected total payment.

For several fixed effort levels, we apply all the performance measurements and estimate their Measurement Integrity and the total payment needed to elicit that equilibrium with the Borda-count payment scheme by iterating the data-generating process 5000 times.

Recall that to satisfy Individual Rationality, the total payment needs to be at least the agents' cost to exert the effort. Since the minimal payment is the same for all performance measurements

[2]Note that even though DMI demonstrates impressive theoretical properties, it perform badly in our simulation due to the considerable noise of its score (Figure 5). Thus, we do not show it in the other figures.
[3]Note that the total payment of Borda-count is deterministic, thus, we use "total payment" in the rest of this section instead of "expected total payment".

when the effort level is the same, we visualize that as a horizontal line in our result.

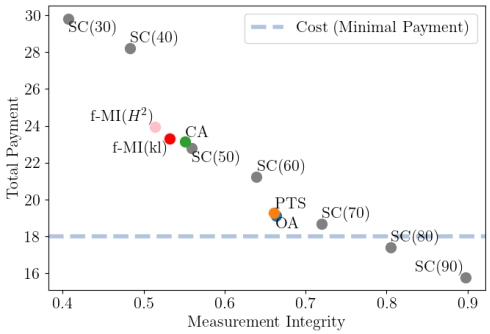


Figure 3: Measurement Integrity v.s Total Payment of Borda-count payment scheme at effort level 0.6.[4] The x-axis is the Measurement Integrity and the y-axis is the total payment needed to elicit that equilibrium within the tournament payment scheme. The horizontal line shows the agents' cost to exert the effort level, which implies the minimal payment to satisfy the Individual Rationality.

With the results in Figure 3, we can observe that:

(1) The Measurement Integrity monotonically increases with the spot-checking ratio.
(2) The Measurement Integrity and the total payment are significantly negatively correlated.

This implies that, at a symmetric equilibrium, if a performance measurement M has Measurement Integrity equal to spot-checking *SCE%*, then it has a similar motivational proficiency, i.e. similar total payment, to that spot-checking performance measurement within a tournament payment scheme. Therefore, a higher Spot Check Equivalence indicates higher motivational proficiency.

Note that, even if considering IR, the motivational proficiency is still monotonically increasing with the Spot Check Equivalence, however, when IR is binding, more spot-checking does not further decrease the total payment.

5.2.1 Measurement Integrity is a computationally efficient proxy. In addition, we find that the Measurement Integrity converges significantly faster than the total payment. The detailed result is demonstrated in Appendix D and Figure 7. This suggests that even when it's possible to compute the expected total payment (e.g. with agent-based model simulation), utilizing Measurement Integrity as a proxy offers better computational efficiency.

5.3 Sensitivity-based Spot Check Equivalence

We now examine the effectiveness of the Spot Check Equivalence based on Sensitivity as a metric of motivational proficiency. Here, when estimating $\Delta\mu/\sigma$ (which is proportional to the Sensitivity), we consider a more realistic scenario where there is only one sample from the data generating process and no ground truth. We estimate $\Delta\mu/\sigma$ for each performance measurement according to Algorithm 1 with $T = 5000$ iterations. We then compare the $\Delta\mu/\sigma$ with the total payment estimated in the same way as in the previous subsection. The results are shown in Figure 4.

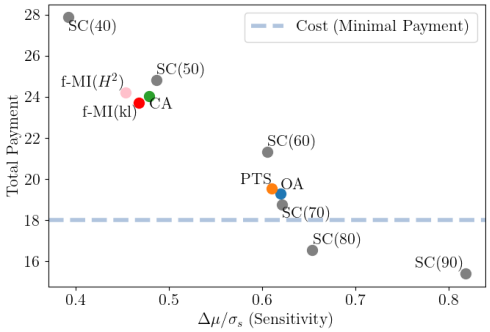


Figure 4: Sensitivity v.s Total Payment of Borda-count payment scheme at effort level 0.6.[4] The x-axis is the $\Delta\mu/\sigma$ and the y-axis is the total payment needed to elicit that equilibrium within the tournament payment scheme.

Similarly, with the results in Figure 4, we can observe that: (1) The Sensitivity monotonically increases with the spot-checking ratio. (2) The Sensitivity and the total payment are significantly negatively correlated. This observation implies that the Spot Check Equivalence based on Sensitivity can be used as a metric of motivational proficiency

5.4 Spot Check Equivalence of Peer Prediction

5.4.1 Peer Prediction works better to elicit high effort. We apply the workflow in our agent-based model experiment to further compare the Spot Check Equivalence of different performance measurements in various contexts. Previous works [2, 6, 31] conduct comparisons of certain metrics in specific contexts. To further study the motivational proficiency of the performance measurements, we calculate the Spot Check Equivalence across various information elicitation contexts.

We then enumerate the effort levels and calculate the Spot Check Equivalence via the Measurement Integrity of each performance measurement (Figure 5).

We find that when eliciting a low-effort equilibrium the Spot-Checking Equivalence of peer prediction performance metrics is low, however, the relative motivational proficiency of peer prediction increases fast. That is because a peer prediction performance measurement scores an agent according to the correlation between her report and her peers', when every agent exerts a low effort, her peers' reports are noisy so the score is quite noisy.

5.4.2 Mutual-information-based mechanisms work better when #tasks per agent increases. As the number of tasks per agent increases, the SCE of OA and PTS remains the same, while for the mutual-information-based mechanisms (CA, f-MI), the Spot Check Equivalence significantly increases. As the #tasks per agent increases,

[4]The figures for other effort levels are shown in our arXiv version.

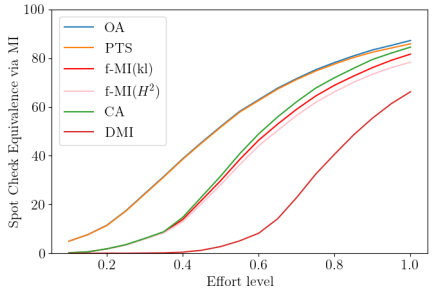
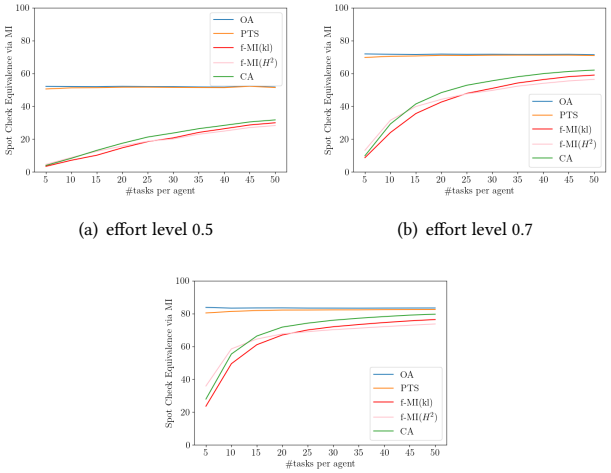


Figure 5: Spot Check Equivalence on different effort levels

the estimation of the joint distribution of two agents' reports will be more accurate, which leads to a more accurate score.



(a) effort level 0.5

(b) effort level 0.7

(c) effort level 0.9

Figure 6: #tasks per agent vs. the Spot Check Equivalence based on Measurement Integrity: the x-axis is the number of tasks assigned to each agent and the y-axis is the Spot Check Equivalence calculated by Measurement Integrity.

6 DOES PEER PREDICTION MAKE THINGS WORSE?

We now delve into the results of Gao et al. [6], especially their assertion that "peer prediction makes things worse", a claim which contradicts other studies. As we discussed in Section 1, their results rely on some restrictive assumptions.

First, they assume that payments are additive across tasks. In contrast, Zhang and Schoenebeck [31] uses tournaments, which are not linear and typically have much better motivational proficiency.

Second, they assume a fixed payment function f for each task. However, as we have seen, a decrease in the spot-checking ratio can be offset by scaling payments.

Finally, the model of Gao et al. [6] assumes "cheap signals", which are signals that agents can coordinate on more easily than the signal the mechanism really desires. For example, when peer reviewing an article, it is much easier to assess quality from the author list than by meticulously reading the article to asses the quality of its argument. By reporting this "cheap signal" instead of the information desired by the mechanism agents can coordinate more with less effort and defeat certain peer-prediction mechanisms.

In response to this analysis, subsequent research has proposed peer-prediction mechanisms that aim to be robust against "cheap" signals Kong and Schoenebeck [14]. The essential idea is to ask the agents to additionally provide the cheap signals, and then pay them for coordination in addition to these signals.

A more thorough explanation is provided in Appendix E.

7 RELATED WORK

Besides the theoretical literature discussing peer prediction mechanisms, as highlighted in Section 5, there are empirical studies that validate these mechanisms. For instance, Radanovic et al. [21] experimentally tested their peer prediction mechanism in both peer grading and crowdsourcing scenarios to validate its theoretical properties. Shnayder et al. [24] employed peer grading data from the edX MOOC platform to assess the performance of their proposed mechanisms. Spot checking in peer grading scenarios has already been empirically examined in works such as [29, 30]. Additionally, Goel and Faltings [8] study combining peer prediction and spot-checking, and introduce the Deep Bayesian Trust Mechanism that utilizes peer reports to reduce the need for spot-checking.

Furthermore, in forecasting contexts where agents are rewarded afterward based on the agreement between their forecasts and the outcomes, i.e. the ground truth is accessible for little or no cost, Hartline et al. [9], Li et al. [18], Neyman et al. [20] study how to optimize proper scoring rules to incentivize effort, and consequently, elicit high-quality information. These works suggest the possibility of optimizing the spot-checking mechanisms by scoring the checked tasks according to the optimal proper score rules, which indicates possible direction for future research.

8 CONCLUSION AND DISCUSSION

In summary, our research provides a methodology for understanding the performance, especially motivational proficiency, of information elicitation mechanisms in various contexts, the Spot Check Equivalence, and consequently offers valuable insights for the design of effective and efficient incentive mechanisms that promote the acquisition of high-quality information.

Future research might be conducted to investigate motivational proficiency in a non-monetary setting, e.g. in peer grading, we care about how to elicit agents' effort with the bounded individual payoff, since the students' grades could only be A, B, C, F, etc. Another future direction might be to study motivational proficiency in a more sophisticated model where the agents have heterogeneous cost functions.

ACKNOWLEDGMENTS

We would like to thank Noah Burrell for his invaluable help and profound discussions that are instrumental in initiating this paper.

REFERENCES

[1] Arpit Agarwal, Debmalya Mandal, David C Parkes, and Nisarg Shah. 2017. Peer Prediction with Heterogeneous Users. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, 81–98.

[2] Noah Burrell and Grant Schoenebeck. 2021. Measurement Integrity in Peer Prediction: A Peer Assessment Case Study. *arXiv preprint arXiv:2108.05521* (2021).

[3] Jenny J Chen, Natala J Menezes, Adam D Bradley, and T North. 2011. Opportunities for crowdsourcing research on amazon mechanical turk. *Interfaces* 5, 3 (2011), 1.

[4] Anirban Dasgupta and Arpita Ghosh. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*. 319–330.

[5] Boi Faltings, Radu Jurca, and Goran Radanovic. 2017. Peer truth serum: incentives for crowdsourcing measurements and opinions. *arXiv preprint arXiv:1704.05269* (2017).

[6] Alice Gao, James R Wright, and Kevin Leyton-Brown. 2016. Incentivizing evaluation via limited access to ground truth: Peer-prediction makes things worse. *arXiv preprint arXiv:1606.07042* (2016).

[7] Xi Alice Gao, James R. Wright, and Kevin Leyton-Brown. 2019. Incentivizing evaluation with peer prediction and limited access to ground truth. *Artificial Intelligence* 275 (2019), 618–638. https://doi.org/10.1016/j.artint.2019.03.004

[8] Naman Goel and Boi Faltings. 2019. Deep bayesian trust: A dominant and fair incentive mechanism for crowd. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1996–2003.

[9] Jason D Hartline, Liren Shan, Yingkai Li, and Yifan Wu. 2022. Optimal Scoring Rules for Multi-dimensional Effort. *arXiv preprint arXiv:2211.03302* (2022).

[10] Radu Jurca and Boi Faltings. 2005. Enforcing truthful strategies in incentive compatible reputation mechanisms. In *International Workshop on Internet and Network Economics*. Springer, 268–277.

[11] Yuqing Kong. [n. d.]. *Dominantly Truthful Multi-task Peer Prediction with a Constant Number of Tasks*. 2398–2411. https://doi.org/10.1137/1.9781611975994.147 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611975994.147

[12] Yuqing Kong. 2020. Dominantly truthful multi-task peer prediction with a constant number of tasks. In *Proceedings of the fourteenth annual acm-siam symposium on discrete algorithms*. SIAM, 2398–2411.

[13] Yuqing Kong, Yunqi Li, Yubo Zhang, Zhihuan Huang, and Jinzhao Wu. 2022. Eliciting thinking hierarchy without a prior. *Advances in Neural Information Processing Systems* 35 (2022), 13329–13341.

[14] Yuqing Kong and Grant Schoenebeck. 2018. Eliciting expertise without verification. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 195–212.

[15] Yuqing Kong and Grant Schoenebeck. 2018. Water from two rocks: Maximizing the mutual information. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 177–194.

[16] 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 (TEAC)* 7, 1 (2019), 1–33.

[17] Nathan Lambert. 2023. *Llama 2: an incredible open LLM*. https://www.interconnects.ai/p/llama-2-from-meta?sd=pf

[18] 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*. 988–989.

[19] Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51, 9 (2005), 1359–1373.

[20] 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*. 718–733.

[21] Goran Radanovic, Boi Faltings, and Radu Jurca. 2016. Incentives for effort in crowdsourcing using the peer truth serum. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 4 (2016), 1–28.

[22] Grant Schoenebeck and Fang-Yi Yu. 2020. Learning and strongly truthful multi-task peer prediction: A variational approach. *arXiv preprint arXiv:2009.14730* (2020).

[23] Grant Schoenebeck, Fang-Yi Yu, and Yichi Zhang. 2021. Information Elicitation from Rowdy Crowds. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3974–3986. https://doi.org/10.1145/3442381.3449840

[24] Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C Parkes. 2016. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*. 179–196.

[25] MATTEO VENANZI, William Teacy, Alexander Rogers, and Nicholas Jennings. 2015. Weather Sentiment - Amazon Mechanical Turk dataset. https://eprints.soton.ac.uk/376543/

[26] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 319–326.

[27] Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.

[28] Bo Waggoner and Yiling Chen. 2014. Output agreement mechanisms and common knowledge. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 2. 220–226.

[29] James R Wright, Chris Thornton, and Kevin Leyton-Brown. 2015. Mechanical TA: Partially automated high-stakes peer grading. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*. 96–101.

[30] Hedayat Zarkoob, Hu Fu, and Kevin Leyton-Brown. 2019. Report-sensitive spot-checking in peer-grading systems. *arXiv preprint arXiv:1906.05884* (2019).

[31] Yichi Zhang and Grant Schoenebeck. 2023. High-Effort Crowds: Limited Liability via Tournaments. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3467–3477. https://doi.org/10.1145/3543507.3583334

[32] Yichi Zhang and Grant Schoenebeck. 2023. Multitask Peer Prediction With Task-dependent Strategies. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3436–3446. https://doi.org/10.1145/3543507.3583292

A ILLUSTRATIVE EXAMPLE

We provide a simple example that illustrates the calculation of expected total payment in an information elicitation context with a winner-take-all tournament payment scheme. In the next section, we then demonstrate the Sensitivity and the Measurement Integrity in this example.

Example A.1. *Consider a simple case where there are two agents (1 and 2) working on a large number of tasks. Each agent $i \in \{1, 2\}$ can choose to exert a non-negative effort level e_i, which incurs a cost $c(e_i)$. And we use a quadratic cost function, $c(e_i) = e_i^2$, which is one of the simplest functions that satisfies all the properties of a cost function.*

Since the signals are independent conditioning on an agent's effort level and the spot-checking performance measurement checks each task u.a.r., we use normal distributions to approximate the quality and performance score: Agent i's report quality is $q_i \sim N(e_i, 1)$, and the performance score is $s_i \sim N(q_i, \sigma^2)$, where σ is monotonically decreasing with the spot-checking ratio. Consequently, we have that $s_i \sim N(e_i, \sigma^2 + 1)$

The expected utility of agent i is

$$u_i(e_i) = \Pr[i \text{ wins the tournament}] \cdot \text{payoff} - c(e_i)$$

$\Pr[i \text{ wins the tournament}] = \Pr[s_i - s_{-i} \geq 0]$, *where $-i$ denotes the other agent. Notice that $(s_i - s_{-i}) \sim N(e_i - e_{-i}, 2\sigma^2 + 2)$ because the difference of two normal random variables is also a normal random variable where the variances add. Therefore, we have*

$$\Pr[i \text{ wins the tournament}] = CDF_{N(e_{-i}, 2\sigma^2+2)}(e_i)$$

where $CDF_{N(e_{-i}, 2\sigma^2+2)}$ is the cumulative distribution function of the normal distribution $N(e_{-i}, 2\sigma^2 + 2)$.

To achieve the symmetric local equilibrium at effort level ξ, we let the derivative of agent i's expected utility at $e_i = \xi$ equal to 0.

$$\frac{\partial}{\partial \xi} u_i(\xi) = pdf_{N(\xi, 2\sigma^2+2)}(\xi) \cdot \text{payoff} - 2\xi = 0$$

where $pdf_{N(\xi, 2\sigma^2+2)}$ is the probability density function of the normal distribution $N(\xi, 2\sigma^2 + 2)$, which is the derivative of $\Pr[i \text{ wins the tournament}]$ at $e_i = \xi$. By simplification, we get, $pdf_{N(\xi, 2\sigma^2+2)}(\xi) = \frac{1}{2\sqrt{\pi \cdot (\sigma^2+1)}}$. Solving for the payoff, we have

$$\text{payoff} = 4\xi \cdot \sqrt{\pi \cdot (\sigma^2 + 1)}$$

In addition, we need to keep Individual Rationality, i.e. the expected utility for each agent should not be negative so that rational agents will not leave, which implies the total payoff to all agents should cover the total cost of all agents, i.e., $\text{payoff} \geq 2c(\xi) = 2\xi^2$. *Putting these together, the payoff should be*

$$\text{payoff} = \max\left(2\xi^2, 4\sqrt{\pi \cdot (\sigma^2 + 1)} \cdot \xi\right)$$

In the above example, we note that, σ monotonically decreases with the spot-checking ratio, and the total payoff monotonically increases with σ when IR is not binding. Therefore, we can see that a higher spot-checking ratio leads to a lower total payment when IR is not binding, and consequently, a higher motivational proficiency.

Sensitivity in Example A.1. Recall that Zhang and Schoenebeck [31] propose the Sensitivity, which measures how sensitive the performance score is to the effort change. They show that the total payoff for eliciting effort level ξ is monotonically decreasing with Sensitivity (Lemma 2.7). Here, we show the Sensitivity $\delta(\xi)$ in the above example.

$$\delta(\xi) = \frac{\frac{\partial}{\partial e_i}\mu_s(e_i)|_{e_i=\xi}}{\sigma_s(\xi)} = \frac{1}{\sqrt{\sigma^2 + 1}}$$

Measurement Integrity in Example A.1. In Example A.1, we can see that the motivational proficiency highly depends on how an agent's performance score correlates with his report quality. In the example, the Pearson correlation coefficient between the agent i's report quality and score is as follows.

$$\rho(q_i, s_i) = \frac{\mathbb{E}[q_i \cdot s_i] - \mathbb{E}[q_i]\mathbb{E}[s_i]}{\sigma_{q_i} \cdot \sigma_{s_i}} = \frac{1}{\sqrt{\sigma^2 + 1}}$$

Note that the total payment is inversely proportional to the Pearson correlation coefficient! And the Sensitivity has the same form as the correlation in this example.

This example shows intuitions that the correlation between the agents' report qualities and scores can be a proxy for a performance measurement's motivational proficiency. And both the report quality and score are accessible in real data.

Therefore, we employ Measurement Integrity [2] as our proxy, which measures the expected correlation between the agents' report qualities and the performance scores in a specific model.

B DISCUSSION OF THE TRUTHFUL REPORT ASSUMPTION.

In our Section 2, we assume that the agents will truthfully report their signal. This assumption is reasonable when applying a linear payment scheme given the performance measurement is truthful under certain settings. In particular, Dasgupta and Ghosh [4] propose the first multi-task peer prediction mechanism. Later on, the CA mechanism [24] and f-MI mechanism [16] are proposed to generalize Dasgupta and Ghosh [4]'s mechanism from binary signal space to finite signal space, the MA mechanism [32] generalizes CA to address task-dependent strategies, the DMI mechanism [11] aims to reduce the required number of tasks, Schoenebeck and Yu [22] handle continuous signals with a learning-based mechanism, Agarwal et al. [1] deal with the heterogeneous agents, and Schoenebeck et al. [23] address adversarial attacks in peer prediction.

Furthermore, Burrell and Schoenebeck [2] examine the robustness of the truth-telling equilibrium with agent-based model experiments. However, when applying a non-linear payment scheme, the agents may have the incentive to strategically report their signal, e.g. increasing the variance of their score to get a higher probability of being the winner. Zhang and Schoenebeck [31] propose a truthful winner-take-all payment scheme by adding noise to the agents' score which may hurt the incentive for effort. However, further study needs to be conducted to study the robustness of different performance measurements against strategic reports with other non-linear payment schemes. This gap indicates another potential future direction of our research.

C COMPUTE SPOT CHECK EQUIVALENCE

In this section, we present the detailed algorithm to compute the Spot Check Equivalence when the measure $f(IEC \leftarrow SC(X))$ is monotonic with respect to X.

Algorithm 2: Binary Search algorithm for *SCE*

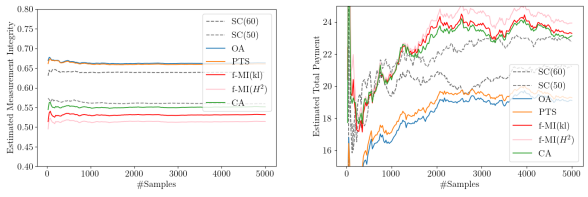
Input: Information Elicitation Context *IEC*, Performance Measurement M, step size ϵ
Output: Spot Check Equivalence *SCE*
Function `BinarySearchSCE`(M, f)**:**
  low = 0, high = $\lfloor 100/\epsilon \rfloor$
  while *low* $\leq$ *high* **do**
    mid = $\left\lfloor \frac{\text{low+high}}{2} \right\rfloor$
    if $f(IEC \leftarrow SC(X = mid * \epsilon)) < f(IEC \leftarrow M)$ **then**
      ans = mid
      low = mid + 1
    end
    else
      high = mid − 1
    end
  end
  return *ans*

Furthermore, a linear combination of the two spot-checking ratios which has adjacent measure f may be applied for a better approximation.

$$SCE = \text{ans} + \epsilon \cdot \frac{f(IEC \leftarrow M) - f(IEC \leftarrow SC(\text{ans}))}{f(IEC \leftarrow SC(\text{ans} + \epsilon)) - f(IEC \leftarrow SC(\text{ans}))}$$

D ADDITIONAL RESULTS

Measurement Integrity is a computationally efficient proxy. Figure 7 illustrates the variation in both the Measurement Integrity and total payment as the number of iterations goes up.



(a) Measurement Integrity (b) Total payment of Borda-count

Figure 7: Convergence speed of the Measurement Integrity and the total payment: the x-axis is the number of the samples, and the y-axis is the estimated Measurement Integrity and the estimated total payment of the Borda-count payment scheme at effort level $\xi = 0.6$ respectively.

In the previous results of Figure 3 (b), we find that when eliciting effort level of $\xi = 0.6$, the f-MI(kl), f-MI(H^2) and CA performance measurements have a little less motivational proficiency comparable to 50% spot-checking. Meanwhile, the OA and PTS performance measurements are better than 60% spot-checking.

Figure 7 demonstrates that achieving the same outcome requires significantly fewer iterations for the calculation of Measurement Integrity compared to the total payment.

E DOES PEER PREDICTION MAKE THINGS WORSE?

In this section, we provide a detailed discussion about why Gao et al. [6, 7] have the result that "peer prediction makes things worse" which contradicts the other literature [31] and our main results that show peer-prediction mechanisms can have non-zero Spot Check Equivalence.

We first introduce the Information Elicitation Context in their paper.

Agent. Agent $= (I, c, \mathbf{e})$. The agent can choose a binary effort $e \in \{0, 1\}$, where exerting high effort has a cost $c(1) = c^E$ and exerting no effort has a cost $c(0) = 0$.

Application Abstraction. In *App* $= (J, \mathcal{GT}, \omega, \Sigma, D)$, they also model signal generation as a random function:

$$D_{signal} : \{0, 1\} \times \mathcal{GT} \to \Delta\Sigma$$

however, importantly, the signals of agents are not i.i.d. across agents. When the agent exerts high effort, she will get a high-quality signal, which is drawn from a distribution conditional on the task's ground truth. When the agent exerts no effort, she will get a low-quality signal that is uncorrelated with the task's ground truth, but, crucially, the no-effort signals are perfectly correlated across no-effort agents—they all receive the same signal.

Performance Measurement. Firstly, they assume that the scoring function in the spot-checking performance measurement (Definition 2.9) can effectively incentivize high effort, i.e.

$$\mathbb{E}\left[S\left(D_{signal}(1, g_j), g_j\right)\right] - c^E > \mathbb{E}\left[S\left(D_{signal}(0, g_j), g_j\right)\right]$$

In addition to the spot-checking performance measurement (Definition 2.9), they propose a spot-checking peer-prediction performance measurement, where for the unchecked tasks, they apply a peer-prediction performance measurement to score the agents.

Payment Scheme. They fix a function $f : \mathcal{GT} \times \Sigma \to \mathbb{R}^*$. The payment scheme is additive across tasks and pays agents according to f for answering spot-checked tasks, and according to a peer-prediction mechanism for tasks that are not spot-checked.

Agent reports. They allow the agents to strategically report their signals.

Equilibrium. They focus on two possible equilibria. In the no-effort equilibrium, each agent exerts no effort and uses the same strategy to report her signal. In the truthful equilibrium, each agent exerts high effort and truthfully reports her signal.

E.1 Comparing Spot-checking Mechanisms and Spot-checking Peer-prediction Mechanisms

Notice that, by assumption, with enough spot-checking, you will get the truthful equilibrium. They use the minimum spot-checking ratio that ensures the truthful strategy profile is an equilibrium as a measure of the mechanisms' performance. Formally, they use p_{Pareto} to denote the minimum spot-checking ratio where the truthful equilibrium Pareto dominates the no-effort equilibrium when applying the hybrid peer-prediction/spot-checking mechanism. They use p_{ds} to denote the minimum spot-checking ratio where the truthful strategy profile is a dominant strategy in spot-checking mechanism.

Then they propose a theorem for comparing the spot-checking mechanism and the spot-checking peer-prediction mechanism:

Theorem E.1 (Section 5 Theorem 3 in Gao et al. [6]). *For any spot-checking peer-prediction mechanism, if the no-effort equilibrium exists and Pareto dominates the truthful equilibrium when the cost of effort is 0 ($c^E = 0$) and no task is checked ($p = 0$), then $p_{Pareto} \geq p_{ds}$ for any $c^E \geq 0$.*

There are three differences between the models in Gao et al. [6] and Zhang and Schoenebeck [31] that account for this stark difference.

The most obvious difference is that the payments in Gao et al. [6] are restricted to be additive across tasks. This is rather similar to assuming a linear payment rule when the score is additive across tasks. However, Zhang and Schoenebeck [31] use tournaments, which are not linear and typically have much better motivational proficiency.

Second, Gao et al. [6]'s assumption of a fixed payment function f is extremely restrictive. As we discussed in Section 1, when the spot-checking ratio decreases, it is possible to maintain the same incentive properties by simply scaling up the payment function. Thus, if we scale up f, we can make p_{ds} arbitrarily small, and, conversely, by scaling down f, we can make p_{ds} arbitrarily close to 1. On the other hand, in Zhang and Schoenebeck [31] the mechanism is defined as a performance measurement and a payment scheme, and the payment scheme is optimized to work with the scoring function, rather than being artificially fixed.

Finally, Gao et al. [6]' make a key assumption on the no-effort signals being perfectly correlated. For example, in peer grading, the writing and formatting quality is a signal that can be accessed with very little effort while assessing the correctness both requires more effort and will likely lead to less agreement.

Notice that the premise of Theorem E.1 is that when the cost of effort is 0 ($c^E = 0$) and no task is checked ($p = 0$), the equilibrium where agents exert no effort Pareto dominates the truthful equilibrium.

Let's zoom in on this. First, consider the case where the cost of the high effort signal $c^E = 0$. Notice that any spot-checking mechanism that checks any positive ratio of tasks will have the high-effort profile as an equilibrium because agents will receive some positive payoff, but have 0 cost.

Next, consider a peer-prediction mechanism (e.g. a hybrid peer-prediction/spot-checking mechanism with spot-checking ratio 0). Again, the theorem is vacuously true, unless the profile in which no agent exerts effort Pareto dominates the high-effort equilibrium because, in the former, all agents agree and receive a maximal payoff[5], but, in the latter, they do not all agree.

Together, this shows that any peer prediction mechanisms will have a Spot Check Equivalence of 0, since in this case, any spot-checking ratio that is greater than 0 will inevitably lead to the truthful equilibrium given that the effort cost is nonexistent.

Because having a Spot Check Equivalence of 0 is an assumption of the theorem, it is no wonder that such peer-prediction mechanisms do not help!

However, this still potentially makes for a very strong critique of peer-prediction mechanisms because in many real-world settings, there exist cheap signals. For example, as mentioned in the introduction, when humans are labeling LLM responses, it is much easier to judge them on how authoritative-sounding the responses are than on how truthful the responses are. However, such labels may encourage hallucinations.

Indeed, Gao et al. [6] led to several papers trying to create peer-prediction mechanisms that are robust against "cheap" signals (i.e. the no-effort/low-effort signals that can bring higher agreement than the high-effort signals).

Kong and Schoenebeck [14] propose a peer prediction mechanism called Hierarchical Mutual Information Paradigm (HMIP), assuming a hierarchical information structure where high-effort (or higher expertise) agents have access to the information of low-effort (or lower expertise) agents. HMIP encourages agents to invest effort and incentivizes truthful reporting by paying the high-effort agents for correctly predicting the "cheap" signals from the low-effort agents. Additionally, a human subject experiment [13] shows evidence of a hierarchical information structure among the participants.

[5] They are assuming here that complete agreement brings a maximum payoff.