Stochastically Dominant Peer Prediction

Yichi Zhang
DIMACS, Rutgers University
yz1636@dimacs.rutgers.edu

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

David Pennock
DIMACS, Rutgers University
dpennock@dimacs.rutgers.edu

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

Abstract

Eliciting reliable human feedback is essential for many machine learning tasks, such as learning from noisy labels and aligning AI systems with human preferences. Peer prediction mechanisms incentivize truthful reporting without ground truth verification by scoring agents based on correlations with peers. Traditional mechanisms, which ensure that truth-telling maximizes the **expected scores** in equilibrium, can elicit honest information while assuming agents' utilities are **linear functions** of their scores. However, in practice, non-linear payment rules are usually preferred, or agents' utilities are inherently non-linear.

We propose *stochastically dominant truthfulness (SD-truthfulness)* as a stronger guarantee: the score distribution of truth-telling stochastically dominates all other strategies, incentivizing truthful reporting for a wide range of monotone utility functions. Our first observation is that no existing peer prediction mechanism naturally satisfies this criterion without strong assumptions. A simple solution—rounding scores into binary lotteries—can enforce SD-truthfulness, but often degrades *sensitivity*, a key property related to fairness and statistical efficiency. We demonstrate how a more careful application of rounding can better preserve sensitivity. Furthermore, we introduce a new enforced agreement (EA) mechanism that is theoretically guaranteed to be SD-truthful in binary-signal settings and, under mild assumptions, empirically achieves the highest sensitivity among all known SD-truthful mechanisms.

1 Introduction

Information elicitation studies how to design mechanisms that incentivize honest and high-effort human feedback. In this framework, a principal seeks reliable responses regarding a set of tasks from strategic agents who are motivated by the designed rewards. For example, AI companies collect human preferences to align language models via RLHF [Ouyang et al., 2022], and MOOCs use peer grading to evaluate complex assignments at scale Piech et al. [2013]. A key challenge is designing scoring mechanisms that accurately assess the informativeness of agents' reports.

Peer prediction provides an elegant solution for settings where ground truth is unavailable or costly to obtain, such as in subjective tasks [Miller et al., 2005]. Instead of relying on external validation, peer prediction scores agents based on the correlation between their reports and those of their peers. The scoring mechanism is designed to be *truthful* so that everyone reporting truthfully forms a Bayesian Nash equilibrium, ensuring that truth-telling maximizes each agent's **expected score**.

A key assumption underlying the effectiveness of existing peer prediction mechanisms is that agents' utility is a **linear function** of their score. This assumption is reasonable in some cases, such as when agents are solely motivated by monetary payments, which can be straightforwardly designed as a

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

linear transformation of the peer prediction scores. However, practical implementations often favor non-linear transformations, such as tournaments, to map scores to payments due to their improved budget efficiency and flexibility [Zhang and Schoenebeck, 2023a]. Moreover, agents' intrinsic utility functions are sometimes inherently non-linear. For instance, student graders care more about their final letter grades (e.g., A, B, C) than the real-valued numerical scores assigned by the mechanism. When agents' utilities are non-linear functions of their scores, truth-telling is no longer guaranteed to be an equilibrium strategy.

Motivated by this limitation, we introduce the concept of *stochastic dominance-truthful (SD-truthful)* peer prediction mechanisms. This stronger notion of truthfulness requires that an agent's **score distribution**, when everyone reports truthfully, first-order stochastically dominates the score distribution under any unilateral deviation. As a result, SD-truthful mechanisms incentivize truth-telling for agents with **any utility function that is monotone increasing with the score**.

1.1 Our Results

Our first observation is that **no existing peer prediction mechanism is naturally SD-truthful under general information structures**, i.e., under arbitrary correlations between agents' signals. We examine three widely studied mechanisms that compute scores as the average of multiple simple discrete scores, including *Output Agreement (OA)*, *Peer Truth Serum (PTS)* [Faltings et al., 2017], and *Correlated Agreement (CA)* [Shnayder et al., 2016]. Mechanisms with complex scoring rules—such as those based on continuous-valued estimates of mutual information Kong and Schoenebeck [2019], Kong [2020], Schoenebeck and Yu [2020]—are unlikely to satisfy SD-truthfulness. Yet even these simple mechanisms fail to guarantee SD-truthfulness without strong assumptions on the underlying information structure.

Our second observation is that **achieving SD-truthfulness is easy in principle but challenging to do well in practice**. Given any truthful mechanism $\mathcal{M}$ with bounded scores $S^{\mathcal{M}} \in [S_{\inf}, S_{\sup}]$, we can define a probability by normalizing the score: $\lambda^{\mathcal{M}} = \frac{S^{\mathcal{M}} - S_{\inf}}{S_{\sup} - S_{\inf}} \in [0, 1]$. We then score agents via a binary lottery: $\tilde{S}^{\mathcal{M}} = 1$ with probability $\lambda^{\mathcal{M}}$ and 0 otherwise. Although this approach, called the *direct rounding reduction*, trivially satisfies SD-truthfulness, it greatly reduces *sensitivity*—an important property related to the efficiency and ex-post fairness of a mechanism. At a high level, sensitivity measures how much a mechanism's score responds to changes in the information contained in an agent's reports, normalized by the standard deviation of the scores.

To overcome this limitation, **we propose the *partition rounding reduction*, which helps partially recover the sensitivity of the original mechanism**. The idea is to divide the questions into K disjoint subsets and apply the original mechanism separately within each subset to compute an individual score. The final score for an agent is then the average of K i.i.d. individual scores after applying direct rounding. Intuitively, this approach improves sensitivity by reducing variance through the aggregation of multiple independent scores.

However, the partition rounding process cannot fully recover the sensitivity of the original mechanisms, implying additional room for more sensitive peer prediction mechanisms. **We introduce a novel mechanism called *enforced agreement (EA)* that is shown to be the most sensitive SD-truthful mechanism in the binary-signal setting**. EA aims to replicate the key advantage of OA—its high sensitivity—while relaxing the strong assumptions on the information structure required for SD-truthfulness.

EA works by enforcing a pre-determined empirical distribution of signals. In the binary setting, if Alice reports 0 on $m_0 > n_0$ questions (where n_0 is the enforced count), the mechanism randomly flips $m_0 - n_0$ of them to 1. This trick removes agents' incentive to over-report majority signals, a strategy that undermines the truthful properties of OA. However, EA is only truthful but not SD-truthful when signals are non-binary, highlighting an important future challenge in designing high-sensitivity SD-truthful mechanisms for more complex settings.

Lastly, our empirical results reinforce and complement our theoretical insights. In the binary-signal setting, we show that even a prior-free implementation of EA—simply enforcing a uniform distribution—outperforms other SD-truthful mechanisms in most settings, including those that require prior knowledge. For non-binary-signal settings, the partition rounding reduction of PTS emerges as the most sensitive SD-truthful mechanism. As a robustness check, we validate our findings by

assessing budget efficiency, another practical metric of peer prediction mechanisms that quantifies the minimum budget required to elicit a desired level of effort [Zhang and Schoenebeck, 2023a].

Table 1 summarizes a comparison of SD-truthful (implementations of) peer prediction mechanisms with respect to the assumptions they require and their sensitivities.[1]

Mechanisms	Requirements for SD-truthfulness
Output Agreement (OA)	Self-dominating signal
Enforced Agreement (EA) [This paper]	Binary and self-predicting signal
Peer Truth Serum (PTS) [Faltings et al., 2017]	Self-predicting signal
Correlated Agreement (CA)[Shnayder et al., 2016]	Any signal
Matching Agreement (MA) [Zhang and Schoenebeck, 2023b]	Any signal

Table 1: A Comparison of Discussed Peer Prediction Mechanisms ordered by sensitivity (from high to low), where PTS, CA, and MA require the partition rounding reduction to achieve SD-truthfulness.

2 Related Work

The idea of peer prediction was proposed by Miller et al. [2005]. As the first discussion, their mechanism is not *detail-free*, i.e. the mechanism requires prior knowledge of the information structure (how agents' signals are correlated) to obtain truthfulness. There are two ideas to mitigate this limitation: (i) the single-task mechanism, which elicits the information structure directly from agents [Prelec, 2004]; and (ii) the multi-task mechanism, which assumes each agent responds to multiple i.i.d. tasks but only requires eliciting signals rather than predictions [Dasgupta and Ghosh, 2013].

Our study is primarily relevant to the multi-task setting. We categorize multi-task peer prediction mechanisms into two types: peer agreement mechanisms Dasgupta and Ghosh [2013], Faltings et al. [2017], Shnayder et al. [2016], Zhang and Schoenebeck [2023b], Agarwal et al. [2017] and mutual information mechanisms Kong and Schoenebeck [2019], Kong [2020], Schoenebeck and Yu [2020], Schoenebeck et al. [2021], Kong [2024]. In this paper, we primarily focus on peer agreement mechanisms, as SD-truthfulness requires the mechanism's score to take a relatively simple form. Intuitively, this is because a complex scoring mechanism provides more opportunities for a cheating strategy to obtain one of the highest scores with a larger probability than truth-telling, which breaks SD-truthfulness. In Appendix A, we provide a more detailed discussion on why several classic mutual information mechanisms are not SD-truthful.

The score of a peer agreement mechanism can be viewed as an average of multiple individual scores, where each individual score takes values from a small finite space, e.g. $\{0, 1\}$. For example, output agreement (OA) scores an agent based on the average number of questions that they agree with another agent. The first multi-task peer prediction mechanism lies in this category, which can achieve truthfulness in a binary-signal setting [Dasgupta and Ghosh, 2013]. The correlated agreement (CA) mechanism generalizes their idea to any finite-signal setting [Shnayder et al., 2016]. The peer truth serum (PTS) mechanism is a generalization of OA, which relaxes the self-dominating assumption to self-prediction. In addition, two follow-ups of CA aim to handle heterogeneous agents [Agarwal et al., 2017] and more general cheating strategies [Zhang and Schoenebeck, 2023b].

Moreover, we note that the idea of rounding peer prediction scores into a binary lottery (see Section 5) was previously proposed by Miller et al. [2005] to address risk-averse agents. What is new in our work is the comparative analysis of various rounding reductions based on sensitivity, allowing us to identify a more effective rounding approach. Furthermore, we formally prove that this idea extends beyond risk-averse agents and can be applied to any agents with increasing utility functions.

3 Model

We consider the multi-task peer prediction setting where two agents, Alice and Bob, answer the same n i.i.d. questions. For each question j, the agents receive discrete signals $X_{a,j}, X_{b,j} \in \Sigma = \{0, \dots, c-1\}$, drawn from a fixed joint distribution $\Pr(X_a, X_b)$. In cases with more than two agents,

[1] MA is an extension of CA, and therefore shares similar requirements and properties.

we can reduce to the case of two agents by, for any particular agent "Alice", choosing a random peer agent "Bob" for scoring.

Each agent $i \in \{a, b\}$ applies a strategy $\theta_i \in \Theta$, which maps from signals to distributions over the same space Δ_Σ. The report on question j is denoted as $\hat{X}_{i,j} = \theta_i(X_{i,j})$. We highlight two types of strategies: the truth-telling strategy τ, which always reports the received signal, and the uninformative strategy μ, where reports are independent of the input. Examples of uninformative strategies include always reporting the same signal or randomly reporting signals according to the prior.

A peer prediction mechanism $\mathcal{M}$ maps vectors of reports $\hat{X}_a$ and $\hat{X}_b$ to a score for Alice. Given an information structure (i.e., the joint distribution over signals) and a mechanism, the score is a random function of strategies, denoted as $S^{\mathcal{M}}(\theta_a, \theta_b)$. The randomness arises from the signals, strategies, and the mechanism itself. When it is clear from context, we omit the superscript $\mathcal{M}$.

3.1 Truthful Guarantees

Existing peer prediction mechanisms focus on expected scores, guaranteeing that any deviation from truth-telling will only reduce the expected score. In this sense, truth-telling forms a Bayesian Nash Equilibrium. Since our analysis centers on equilibrium behavior, we assume Bob always tells the truth. Under this assumption, the score assigned by the mechanism reduces to a function of Alice's strategy θ, written as $S^{\mathcal{M}}(\theta)$.

Definition 3.1. A peer prediction mechanism is truthful if $\mathbb{E}[S(\tau)] \geq \mathbb{E}[S(\theta)]$ for any $\theta \in \Theta$.

A truthful mechanism guarantees that if agents have linear utility in their scores, truth-telling is a Bayesian Nash Equilibrium—any deviation yields lower expected utility. However, as we argued earlier, assuming linear utility is often undesirable and, in some cases, unrealistic. Instead, we aim to design truthful mechanisms for a broader and more reasonable class of utility functions—those that are simply increasing in score. This motivation naturally leads to the concept of stochastic dominance.

Definition 3.2. A peer prediciton mechanism is stochastic dominance-truthful (SD-truthful) if the distribution of $S(\tau)$ first-order stochastic dominates (FOSD) the distribution of $S(\theta)$ for any $\theta \in \Theta$, i.e. $\Pr(S(\tau) \geq t) \geq \Pr(S(\theta) \geq t)$ for any $t \in \mathbb{R}$.

A standard result in microeconomic theory suggests that $S(\tau)$ FOSD $S(\theta)$ if and only if $\mathbb{E}[u(S(\tau))] \geq \mathbb{E}[u(S(\theta))]$ for every increasing utility function u [Hadar and Russell, 1969]. Since linear utility functions are also increasing, every SD-truthful mechanism is also truthful.

3.2 Sensitivity

In practice, an effective scoring mechanism should meaningfully evaluate the information contained in agents' reports and reward more informative reports with higher scores. This motivates an additional design criterion, orthogonal to truthfulness, known as *sensitivity* [Zhang and Schoenebeck, 2023a].

To motivate the concept of sensitivity, we extend the model to incorporate agents' effort levels. This extension captures how the mechanism responds to marginal changes in information at equilibrium, i.e. how the score varies when all other agents exert full effort while one agent adjusts her effort. In particular, on each question, we assume Alice exerts effort with probability $e \in [0, 1]$, while Bob always exerts effort and reports truthfully. When Alice exerts effort, the signal pair (X_a, X_b) is drawn from the joint distribution $\Pr(X_a, X_b)$; in the extreme case where Alice exerts no effort, the signals are independently drawn from the marginals $\Pr(X_a)\Pr(X_b)$. That is, exerting effort produces a signal X_a that is correlated with X_b according to the joint distribution, while shirking yields an uninformative guess drawn from the prior $\Pr(X_a)$. After observing the signal, Alice will report truthfully. Then, Alice's score under mechanism $\mathcal{M}$ becomes a function of e, denoted as $S^{\mathcal{M}}(e)$.

Definition 3.3. The sensitivity of a peer prediction mechanism $\mathcal{M}$ at effort level e is $\delta^{\mathcal{M}}(e) := \frac{\nabla \mathbb{E}[S^{\mathcal{M}}(e)]}{\text{std}(S^{\mathcal{M}}(e))}$ where ∇ denotes the derivative operator and std denotes the standard deviation.

Sensitivity measures how well a mechanism's expected score responds to changes in effort—and thus, changes in information. Consequently, a trivial mechanism that always assigns a score of 1, though truthful, has zero sensitivity, as its output is unaffected by effort.

Prior work [Zhang and Schoenebeck, 2023a] shows that when the score is normally distributed and agents are rewarded based on the ranking of scores, sensitivity is a sufficient statistic for *budgetary efficiency*: mechanisms with higher sensitivity can elicit the same equilibrium effort level at a lower budgetary cost. Moreover, Xu et al. [2024] establish a one-to-one correspondence between sensitivity and *measurement integrity* [Burrell and Schoenebeck, 2023], implying that mechanisms with higher sensitivity generate scores more closely aligned with the true quality of responses. Together, these findings highlight sensitivity as a practically meaningful and theoretically grounded property of peer prediction mechanisms, motivating our choice to use it as an additional dimension for comparison.

Our goal is to design SD-truthful mechanisms that also achieve high sensitivity.

4 Prior Mechanisms

To the best of our knowledge, no existing peer prediction mechanism naturally achieves SD-truthfulness without imposing strong assumptions on the information structure. In this section, we introduce three exemplary multi-task peer prediction mechanisms and explain why they are not naturally SD-truthful to motivate our methods.

4.1 Output Agreement (OA)

The *output agreement (OA)* mechanism scores Alice based on the fraction of questions on which both agents agree. The final score of OA is the average of n independent binary individual scores, i.e. $S^{OA} = \frac{1}{n}\sum_{i\in[n]} S_i^{OA}$ where $S_i^{OA} = 1$ if both agents' reports agree on question i and 0 otherwise. However, OA encourages over-reporting of the majority signal, which is truthful only under a strong assumption.

Definition 4.1. Agents' signals are self-dominating if $\Pr(X_b = \sigma | X_a = \sigma) > \Pr(X_b = \sigma' | X_a = \sigma)$ for any $\sigma' \neq \sigma \in \Sigma$.

Proposition 4.2. *OA is truthful if agents' signals are self-dominating.*

In fact, OA is also SD-truthful when signals are self-dominating. This is because each individual score S_i^{OA} takes only two values, meaning that $S_i^{OA}(\tau)$ FOSD $S_i^{OA}(\theta)$ if and only if $\Pr(S_i^{OA}(\tau) = 1) \geq \Pr(S_i^{OA}(\theta) = 1)$. Furthermore, the average of multiple i.i.d. scores that are each SD-truthful remains SD-truthful (see Section 5.2). Consequently, for OA, SD-truthfulness reduces to truthfulness.

4.2 Peer Truth Serum (PTS)

Self-dominance is usually considered a strong assumption. [Radanovic et al., 2016] propose the *Peer Truth Serum (PTS)* which incentivizes truth-telling under a weaker assumption.[2]

Definition 4.3. Agents' signals are self-predicting if $\Pr(X_b = \sigma | X_a = \sigma) > \Pr(X_b = \sigma | X_a = \sigma')$ for any $\sigma' \neq \sigma \in \Sigma$.

PTS assumes a symmetric, publicly known prior $R(\sigma) = \Pr(X_a = \sigma) = \Pr(X_b = \sigma)$. It repeatedly samples questions without replacement, and assigns an individual score $S_i^{PTS} = 1/R(\sigma)$ to Alice if $\hat{X}_{a,i} = \hat{X}_{b,i} = \sigma$, and 0 otherwise. The final score of PTS is the average of all S_i^{PTS}.

Proposition 4.4 ([Faltings et al., 2017])**.** *PTS is truthful if agents' signals are self-predicting.*

Is PTS SD-truthful? The short answer is no. Each individual score S_i^{PTS} takes at most $|\Sigma| + 1$ values, with the maximum being $S_{\max} = 1/R(\sigma_{\min})$, where $\sigma_{\min}$ is the least likely signal under the prior. A necessary condition for PTS to be SD-truthful is that truth-telling maximizes $\Pr\left(S_i^{PTS}(\tau) = S_{max}\right)$, which is the probability of Alice agreeing with Bob on signal σ_{min}. While truth-telling, this probability is the joint distribution $\Pr(X_a = \sigma_{min}, X_b = \sigma_{min})$. However, if Alice uses an uninformative strategy μ that always reports $\sigma_{\min}$, this probability increases to $\Pr(X_b = \sigma_{\min})$. Therefore, although μ decreases the expected score, its distribution is not stochastically dominated by truth-telling, meaning that there exists an increasing utility function under which Alice strictly prefers μ to τ.

[2] In the binary signal setting, self-prediction is equivalent to positive correlation, which is a strictly weaker condition than self-dominance.

4.3 Correlated Agreement (CA)

The correlated agreement (CA) mechanism [Shnayder et al., 2016] takes two questions to compute an individual score which has three possible values: -1, 0, and 1. Suppose the joint distribution $\Pr(X_a, X_b)$ is known. The CA mechanism first computes the delta matrix $\Delta_{\sigma,\sigma'} = \Pr(X_a = \sigma, X_b = \sigma') - \Pr(X_a = \sigma)\Pr(X_b = \sigma')$, which is the difference between the joint distribution and the product of marginal distributions. Let $T_\Delta(\sigma, \sigma') = \text{Sign}(\Delta)_{\sigma,\sigma'}$ be the agreement function where $\text{Sign}(x) = 1$ if $x > 0$ and 0 otherwise. Intuitively, $T_\Delta(\sigma, \sigma') = 1$ if the pair of signals (σ, σ') appears more often on the same question than on distinct questions.

Then, CA repeatedly samples a bonus question j and a penalty question k. For an individual score S_i^{CA}, Alice gets 1 if both agents agree on j and gets -1 if her response on j agrees with Bob's response on k, i.e. $S_i^{CA} = T_\Delta(\hat{X}_{a,j}, \hat{X}_{b,j}) - T_\Delta(\hat{X}_{a,j}, \hat{X}_{b,k})$. Since each bonus question can be paired with $n-1$ penalties, the final score is the average over $n(n-1)$ individual scores.

Intuitively, the CA mechanism works by encouraging correlations on the same question and penalizing correlations on distinct questions. Consequently, if Alice always reports the majority signal, the frequency of agreements on the bonus question will be identical to the frequency of agreements on the penalty questions, which results in a score of 0. Prior work shows that when $\Pr(X_a, X_b)$ is known (or accurately estimated), CA is truthful—and in fact, *informed truthful* [Shnayder et al., 2016].

Is CA SD-truthful? The answer depends on the implementation. The following example illustrates that the original implementation of CA described above is not SD-truthful.

Example 4.1. Suppose $n = 2$, so that Alice knows each of the two questions she answers will be used once as the bonus question and once as the penalty question. Her final score is the average of two individual scores, each with $S_i^{CA} \in \{-1, 0, 1\}$. Thus, the final score has five possible values: $S^{CA} \in \{-1, -0.5, 0, 0.5, 1\}$. A necessary condition for SD-truthfulness is that $\Pr(S^{CA}(\tau) = -1) \leq \Pr(S^{CA}(\theta) = -1)$ for any strategy θ. However, we show that this does not hold in general.

Suppose the signals are positively correlated, i.e. T_Δ is a diagonal matrix. Alice receives a score of -1 if and only if she reports different signals on the two questions and Bob disagrees on both. Under truth-telling, this happens with probability $\Pr(S^{CA}(\tau) = -1) = 2\Pr(X_a = 0, X_b = 1)\Pr(X_a = 1, X_b = 0)$. However, if Alice always reports the same signal (e.g., $\mu(\sigma) = 0$ for any σ), she can avoid this outcome entirely. Therefore, μ is not stochastically dominated by τ.

This warns that **repeatedly using the same question as the bonus or the penalty question will sabotage the SD-truthfulness of CA.**

5 A Rounding Reduction For Stochastically Dominant-Truthfulness

We introduce a straightforward rounding reduction that maps any truthful mechanism to an SD-truthful one. However, this approach often comes at a huge cost of sensitivity. To address this, we propose a more refined rounding method that partially restores the original mechanism's sensitivity.

5.1 Direct Rounding

Consider a peer prediction mechanism $\mathcal{M}$ that assigns Alice a score $S^{\mathcal{M}}$ with bounded support $S^{\mathcal{M}}_{\sup} < \infty$ and $S^{\mathcal{M}}_{\inf} > -\infty$. In *direct rounding*, we normalize the score to a probability: $\lambda^{\mathcal{M}} = \frac{S^{\mathcal{M}} - S^{\mathcal{M}}_{\inf}}{S^{\mathcal{M}}_{\sup} - S^{\mathcal{M}}_{\inf}} \in [0, 1]$. Alice's final score $\tilde{S}^{\mathcal{M}}$ is then determined by a binary lottery: receiving 1 with probability $\lambda^{\mathcal{M}}$ and 0 otherwise.

Proposition 5.1. *The direct rounding reduction of a truthful mechanism with bounded scores is SD-truthful.*

Intuitively, because direct rounding always outputs a Bernoulli score, SD-truthfulness simply implies that truth-telling should maximize the expected success probability. But the expected success probability is just the success probability. By our design of $\lambda^{\mathcal{M}}$, the success probability is a linear function of the final score $S^{\mathcal{M}}$, which proves equivalence between SD-truthfulness and truthfulness.

However, scoring agents via a single binary score is clearly not practical. The following proposition suggests that direct rounding usually destroys the sensitivity of the original mechanism. Let $\delta^{\mathcal{M}}(e)$

and $\tilde{\delta}^{\mathcal{M}}(e)$ be the sensitivity of $\mathcal{M}$ and its direct rounding reduction. Additionally, let $m^{\mathcal{M}}(e)$ and $\text{std}^{\mathcal{M}}(e)$ be the expected score and the standard deviation.

Proposition 5.2. *The sensitivity ratio between a mechanism and its direct rounding is:*

$$\frac{\delta^{\mathcal{M}}(e)}{\tilde{\delta}^{\mathcal{M}}(e)} = \frac{\sqrt{\left(S_{\sup}^{\mathcal{M}} - m^{\mathcal{M}}(e)\right)\left(m^{\mathcal{M}}(e) - S_{\inf}^{\mathcal{M}}\right)}}{\mathit{std}^{\mathcal{M}}(e)}.$$

This ratio is typically large. For example, the final score of OA is an averaged Binomial $\text{Bin}(n, m^{OA}(e))$, whose standard deviation is $\sqrt{m^{OA}(e)(1-m^{OA}(e))/n}$. Thus, the sensitivity ratio between OA and its direct rounding is $\sqrt{n}$. This illustrates a key drawback of direct rounding: it greatly reduces the sensitivity by eliminating the benefit of averaging over multiple questions.

5.2 Partition Rounding

To reduce variance of and improve sensitivity, we aim to better leverage the information from all n questions. In particular, if the final score of $S^{\mathcal{M}}$ is the average of K independent individual scores $S_i^{\mathcal{M}}$, we can apply direct rounding to each $S_i^{\mathcal{M}}$ and take the average. This motivates the idea of *partition rounding*. We first partition the n questions into K disjoint subsets, using each to compute an individual score. Each score is then directly rounded as in Section 5.1, and the final score $\hat{S}^{\mathcal{M}}$ is the average of the K rounded scores. The following lemma proves the feasibility of this idea.

Lemma 5.3. *Let $S = \frac{1}{K}\sum_{i\in[K]} S_i$, where the individual scores S_i are i.i.d.. Then, $S(\tau)$ first-order stochastically dominates $S(\theta)$ if and only if each individual score $S_i(\tau)$ first-order stochastically dominates $S_i(\theta)$.*

Proposition 5.4. *The partition rounding reduction of a truthful mechanism with bounded scores is SD-truthful.*

We defer the proof of Lemma 5.3, which follows from a straightforward coupling argument, to Appendix F.3. Proposition 5.4 then follows directly from this lemma and the SD-truthfulness of direct rounding.

We now show that the partition rounding reduction can partially recover the sensitivity of the original mechanism. The key observation is that the final score under partition rounding is the average of a binomial variable, i.e. $\hat{S}^{\mathcal{M}}(e) = \frac{1}{K}\text{Bin}(K, \lambda^{\mathcal{M}}(e))$, where $\lambda^{\mathcal{M}}(e) = \frac{m^{\mathcal{M}} - S_{\inf}^{\mathcal{M}}}{S_{\sup}^{\mathcal{M}} - S_{\inf}^{\mathcal{M}}}$ with $m^{\mathcal{M}}$ being the expectation of the individual score. Putting this into Definition 3.3 gives us the sensitivity of the partition rounding reduction

$$\hat{\delta}^{\mathcal{M}}(e) = \frac{\nabla m^{\mathcal{M}}(e) \cdot \sqrt{K}}{\sqrt{\left(m^{\mathcal{M}}(e) - S_{\inf}^{\mathcal{M}}\right)\left(S_{\sup}^{\mathcal{M}} - m^{\mathcal{M}}(e)\right)}}. \tag{1}$$

The sensitivity of the partition rounding reduction is exactly $\sqrt{K}$ times that of the direct rounding reduction. This means that by using partition rounding, we can recover a substantial portion of the original mechanism's sensitivity.

In Appendix B, we present the sensitivities of the partition-rounded versions of the three mechanisms discussed. We summarize the limitations of these mechanisms below which motivates the design of a new SD-truthful mechanism introduced in the next section.

- OA is already SD-truthful without rounding and thus has a high sensitivity. However, OA is SD-truthful only for self-dominating signals.
- PTS requires rounding to ensure SD-truthfulness. Therefore, when the prior of signals is biased, $S_{\sup}^{PTS}$ becomes large, which in turn leads to a low sensitivity.
- CA uses two questions to compute an individual score, so $K = n/2$. Furthermore, only bonus questions contribute to sensitivity, while penalty questions are used to enforce truthfulness. This dilutes the sensitivity by half again. Therefore, only $1/4$ of the questions count for the sensitivity of the partition-rounded CA. Hence, even without rounding, the sensitivity of the partition-rounded CA is only half of that of the original implementation.

6 The Enforced Agreement Mechanism

This section aims to present a novel mechanism with high sensitivity, which does not require the above rounding reduction. To motivate our idea, we begin by summarizing the key lessons learned from the previous discussions.

First, Example 4.1 shows a particular type of strategy that prevents previous mechanisms from being SD-truthful: reducing/changing the variance of scores, e.g. by always reporting the same signal. This inspires our *enforced agreement (EA)* mechanism, where the idea is to control a randomness level within agents' reports so that strategic manipulations cannot arbitrarily alter the variance of the scores.

Second, our analysis of sensitivity in Section 5.2 suggests that a high-sensitivity mechanism requires a larger number of partitions K and a smaller score range $S_{\sup} - S_{\inf}$. This insight guides us to use just one question to compute an individual score (so $K = n$) and avoid the need for rounding. We show that EA meets these criteria.

We first prove the SD-truthfulness of EA in the binary-signal setting, and then establish a negative result explaining why it fails to ensure SD-truthfulness when signals are non-binary. Nonetheless, we show that EA remains truthful in non-binary settings when the signal structure is known or can be accurately learned.

6.1 EA in the Binary-signal Setting

We introduce EA in the binary-signal setting and defer the discussion for general settings to the appendix. The idea of EA is to control the marginal distribution of Alice's reports. In the binary setting, the mechanism commits to a target empirical distribution $\Phi = (n_0, n_1)$ where $n_0 + n_1 = n$. The mechanism expects Alice to report signal i on exactly n_i out of n questions. If Alice's actual distribution $\Phi_{X_a} = (m_0, m_1)$ and suppose W.L.O.G. that $m_0 > n_0$, the mechanism randomly selects $m_0 - n_0$ of her 0-reports and flips them to 1. After enforcing Φ, Alice is scored using the output agreement mechanism (OA).

Compared with OA, the enforcement process in EA prevents Alice from benefiting by over-reporting the majority signal. However, this enforcement introduces dependencies across questions, so the binary agreement scores are no longer i.i.d., and Lemma 5.3 does not apply. Our main result shows that EA is SD-truthful when signals are self-prediction (Definition 4.3).

Theorem 6.1. *If $|\Sigma| = 2$ and signals are self-predicting, the enforced agreement mechanism is SD-truthful for any Φ.*

We defer the detailed proof to the appendix and present the main ideas below. Note that the final score of the enforced agreement mechanism is the average of n Bernoulli variables, each with a potentially different success probability. In the binary-signal setting, there are four possible success probabilities in total—$p_{ij} = \Pr(X_b = j \mid X_a = i)$ for $i, j \in \{0, 1\}$—representing the probability of Bob observing (and reporting) j on a question conditioned on Alice observing i. A Bernoulli variable with success probability p_{ij} corresponds to a question on which Alice observes i but is flipped to j either by the mechanism or by Alice herself. Suppose W.L.O.G. that $m_0 > n_0$, meaning that the mechanism will randomly flip $m_0 - n_0$ reports of Alice from 0 to 1 under truth-telling. In this case, the final score of Alice is the average of three binomials:

$$n \cdot S^{EA}(\tau) \sim \mathrm{Bin}(n_0, p_{00}) + \mathrm{Bin}(m_1, p_{11}) + \mathrm{Bin}(m_0 - n_0, p_{01}). \tag{2}$$

We show that for any untruthful strategy, the resulting final score can be written as:

$$n \cdot S^{EA}(\theta) \sim \mathrm{Bin}(n_0 - k, p_{00}) + \mathrm{Bin}(m_1 - k, p_{11}) + \mathrm{Bin}(m_0 - n_0 + k, p_{01}) + \mathrm{Bin}(k, p_{10}),$$

where $0 \le k \le \min(n_0, m_1)$. Under the self-prediction condition, $p_{00} > p_{10}$ and $p_{11} > p_{01}$. Therefore, any untruthful strategy can only reallocate k samples from two binomial distributions with higher success probabilities to those with smaller success probabilities. Via a simple coupling argument, we can see that $S(\theta)$ is first-ordered stochastically dominated by $S(\tau)$.

Additional results are provided in Appendix C. First, we show that EA is not generally SD-truthful in the non-binary setting. Strategic signal permutations can produce score distributions that are not stochastically dominated by truth-telling. However, EA remains truthful when the information structure is known or well estimated. Moreover, we derive a closed-form expression for the sensitivity-maximizing Φ, enabling computation of the optimal enforcement from the information structure.

7 Empirical Evaluations of SD-Truthful Mechanisms

In this section, we empirically compare the sensitivity of the proposed mechanisms across various settings. The results support our theoretical findings: EA consistently exhibits the highest sensitivity among SD-truthful mechanisms in the binary-signal setting (except when the signal prior is nearly uniform). We further evaluate each mechanism's budgetary efficiency in an effort-elicitation scenario—measuring the minimum payment required to incentivize a target effort level. Again, EA proves its superiority, which has the best performance against any other SD-truthful and non-SD-truthful mechanisms.

7.1 Datasets and Experiment Setup

We use two real-world datasets to estimate the information structure between two agents. The first dataset contains binary labels classifying whether a compound is appropriate or inappropriate to be synthesized [Baba et al., 2018]. The second dataset collects the annotations of the sentiment of 300 tweets, where the size of the signal space is 4 [Venanzi et al., 2015]. We denote these two information structures as J_1 and J_2 respectively, and defer the details to Appendix E.1.

In the main body of the paper, we focus on SD-truthful mechanisms, which include:

- **OA** (Section 4.1), which is SD-truthful only when signals are self-dominating.
- **PTS-partition-round**—the partition rounding reduction of PTS (Section 4.2), where we re-weight every individual score using signal prior and take the average.
- **CA-partition-round**—the partition rounding reduction of CA (Section 4.3), where we create $K = n/2$ partitions of questions and score agents the average of the individual CA score for each partition after rounding.
- **MA-partition-round**—the partition rounding reduction of the matching agreement mechanism (Appendix D).
- **EA-prior**—EA with enforcement $\Phi = n \cdot \Pr(X_a)$.
- **EA-uniform**—EA with $\Phi_i = n \cdot 1/|\Sigma|,\ \forall i \in \Sigma$.
- **EA-optimal**—EA with the sensitivity-maximizing Φ computed according to Appendix C.3.

For information structures with binary-signal settings, sensitivities can be computed analytically as shown in Appendix B. For general cases, we estimate sensitivity using a Monte Carlo approach. In particular, we simulate reports to each of the n questions with both agents exerting full effort $e = 1$ and run $\mathcal{M}$ to compute the scores. Repeating this process for $T = 20{,}000$ times yields T i.i.d. samples of $S^{\mathcal{M}}(e)$. We then compute the score of Alice when she deviates to a lower effort level at $e - \Delta e = 0.8$, and obtain T samples of $S^{\mathcal{M}}(e - \Delta e)$. The sensitivity of at effort e is then estimated as $\frac{\mathbb{E}[S^{\mathcal{M}}(e)] - \mathbb{E}[S^{\mathcal{M}}(e-\Delta e)]}{\Delta e} \cdot \frac{1}{\mathrm{std}(S^{\mathcal{M}}(e))}$.

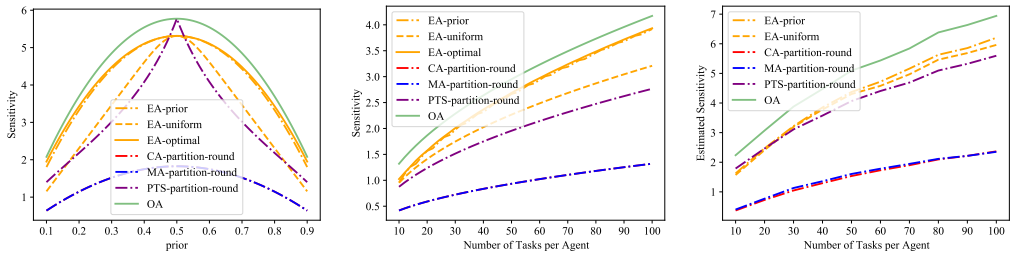
7.2 The Sensitivity of SD-Truthful Mechanisms

We now compare the sensitivity of the discussed SD-truthful mechanisms. In Figure 1a, we show the sensitivity when the prior of the binary signal is varied, i.e. $\Pr[X_a = 0] \in [0.1, 0.9]$, and the number of questions $n = 100$.

Figure 1b and 1c present sensitivity as the number of questions $n \in \{10, 20, \ldots, 100\}$ increases, under information structures J_1 and J_2, respectively. In J_2, where $|\Sigma| > 2$, EA is no longer SD-truthful; however, we still include its sensitivity for comparison.

We make the following observations. First, although the partition reductions of CA and MA are SD-truthful for general settings, their sensitivities are dominated by other SD-truthful mechanisms. This is due to their requirements of the penalty questions, which do not contribute to the sensitivity, and the fact that they require two questions to compute an individual score.

Second, PTS performs well when the signal prior is nearly uniform, i.e. $\Pr(X_a = 0) \approx 0.5$. Intuitively, this is because the rounding factor that enlarges the gap $S^{PTS}_{\sup} - S^{PTS}_{\inf}$ shrinks when the prior is close to uniform. However, even a slight bias, i.e. $|\Pr(X_a = 0) - 0.5| > 0.03$ causes PTS-partition-round to be outperformed by EA.



(a) Varying prior $\Pr[X_a = 0]$ (b) Varying #tasks under J_1 (c) Varying #tasks under J_2

Figure 1: A Comparison of the Sensitivity of Different Mechanisms.

Third, we observe that enforcing the signal prior in EA closely approximates the optimal enforcement. Remarkably, even EA-uniform, a detail-free version that doesn't require knowledge of the prior, consistently outperforms PTS-partition-round, which does rely on prior knowledge for SD-truthfulness.

Additionally, we note that the partition-rounding reduction of MA coincides with that of CA when signals are binary (see Appendix D). Consequently, the blue and red overlap in Figure 1a and 1b.

Due to space constraints, we defer (i) the comparison between partition-rounding reductions and their original implementations to Appendix E.2, and (ii) the details of our effort-elicitation experiments and corresponding budgetary efficiency results to Appendix E.3. These results further show that high-sensitivity mechanisms perform strongly in practical, incentive-aligned environments, underscoring the importance of EA.

8 Conclusion and Future Work

This paper initiates the discussions of stochastically dominant-truthful (SD-truthful) peer prediction mechanisms, which ensure that in equilibrium, the score distribution under truth-telling first-order stochastically dominates any other strategy. Unlike traditional peer prediction mechanisms, which rely on linear utility assumptions, SD-truthful mechanisms can preserve truthfulness for any increasing utility function. We show that existing mechanisms fail to naturally achieve SD-truthfulness and propose a rounding method to enforce it. However, this process often reduces the sensitivity of the original mechanism, making it less efficient in practice. Additionally, we introduce the Enforcement Agreement (EA) mechanism, which is empirically shown to have the highest sensitivity among all SD-truthful mechanisms in the binary-signal setting—except when the signal prior is nearly uniform.

Future work is needed to investigate more sensitive SD-truthful mechanisms beyond the binary-signal setting. A promising next step is to explore approximate SD-truthfulness, which allows truth-telling to lose a small utility for some utility functions. This relaxation could lead to mechanisms with better sensitivity and budget efficiency. Furthermore, SD-truthfulness can be extended to design manipulation-resistant evaluation metrics for AI benchmarking. In ranking-based settings such as Chatbot Arena [Chiang et al., 2024], where models are compared through pairwise judgments, an SD-truthful metric ensures that a model can improve its ranking only by improving its true expected performance, not by exploiting evaluation biases or strategic manipulation.

Broader Impact

On the positive side, our method enhances the reliability and efficiency of crowdsourced data collection, which supports more trustworthy machine learning systems and reduces reliance on expert annotation. However, mechanisms that incentivize effort may unintentionally disadvantage contributors with limited time, resources, or familiarity with the task, raising fairness concerns. We aim to mitigate these risks by designing robust, broadly applicable mechanisms and promoting transparency in their deployment.

Acknowledgment

This work was partly carried out at the DIMACS Center at Rutgers University and supported by NSF #2313137.

References

Arpit Agarwal, Debmalya Mandal, David C Parkes, and Nisarg Shah. Peer prediction with heterogeneous users. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 81–98. ACM, June 2017.

Yukino Baba, Tetsu Isomura, and Hisashi Kashima. Wisdom of crowds for synthetic accessibility evaluation. *Journal of Molecular Graphics and Modelling*, 80:217–223, 2018.

Noah Burrell and Grant Schoenebeck. Measurement integrity in peer prediction: A peer assessment case study. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, page 369–389, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701047. doi: 10.1145/3580507.3597744. URL https://doi.org/10.1145/3580507.3597744.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

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

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28 (1):20–28, 1979.

Boi Faltings, Radu Jurca, and Goran Radanovic. Peer truth serum: Incentives for crowdsourcing measurements and opinions, 2017. URL https://arxiv.org/abs/1704.05269.

Josef Hadar and William R. Russell. Rules for ordering uncertain prospects. *The American Economic Review*, 59(1):25–34, 1969. ISSN 00028282. URL http://www.jstor.org/stable/1811090.

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

Yuqing Kong. Dominantly truthful peer prediction mechanisms with a finite number of tasks. *J. ACM*, 71(2), apr 2024. ISSN 0004-5411. doi: 10.1145/3638239. URL https://doi.org/10.1145/3638239.

Yuqing Kong and Grant Schoenebeck. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. *ACM Transactions on Economics and Computation (TEAC)*, 7(1):1–33, 2019.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in moocs. *arXiv preprint arXiv:1307.2579*, 2013.

Drazen Prelec. A bayesian truth serum for subjective data. *science*, 306(5695):462–466, 2004.

Goran Radanovic, Boi Faltings, and Radu Jurca. Incentives for effort in crowdsourcing using the peer truth serum. *ACM Trans. Intell. Syst. Technol.*, 7(4), March 2016. ISSN 2157-6904. doi: 10.1145/2856102. URL https://doi.org/10.1145/2856102.

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

Grant Schoenebeck, Fang-Yi Yu, and Yichi Zhang. Information elicitation from rowdy crowds. In *Proceedings of the Web Conference 2021*, WWW '21, page 3974–3986, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449840. URL https://doi.org/10.1145/3442381.3449840.

Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C. Parkes. Informed truthfulness in multi-task peer prediction, 2016. URL https://arxiv.org/abs/1603.03151.

Matteo Venanzi, William Teacy, Alexander Rogers, and Nicholas Jennings. Weather sentiment - amazon mechanical turk dataset, 2015. URL https://eprints.soton.ac.uk/376543/.

Shengwei Xu, Yichi Zhang, Paul Resnick, and Grant Schoenebeck. Spot check equivalence: an interpretable metric for information elicitation mechanisms. *arXiv preprint arXiv:2402.13567*, 2024.

Yichi Zhang and Grant Schoenebeck. High-effort crowds: Limited liability via tournaments. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3467–3477, New York, NY, USA, 2023a. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583334. URL https://doi.org/10.1145/3543507.3583334.

Yichi Zhang and Grant Schoenebeck. Multitask peer prediction with task-dependent strategies. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3436–3446, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583292. URL https://doi.org/10.1145/3543507.3583292.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please see abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: As mentioned several times in the paper, our solution requires a binary-signal assumption to achieve the strong truthful guarantee proposed in the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification:Please see Section 3 for detailed assumptions and theory. Proofs are left in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The datasets and parameters of our experiments are presented in Section 7

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the code for our experiments in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Our experiments do not require training. Details of set-up and parameters are shown in Section 7 and Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our figures are visualizations of analytical results where there is no error, or averaged using a large number of independent trials so that the error bars are all very close to zero.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Our experiments do not involve dense computing, and can be conducted on laptops.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: There is no violation of the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide a broader impacts section right after the conclusion section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This work does not release any models or datasets with a high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We believe all datasets and LLMs used in this paper are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This paper releases code used for running the experiments. The code is documented and provided alongside the submission to ensure reproducibility. No new datasets or models are introduced; all experiments are based on public or simulated data.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve data collection from human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve data collection from human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were not used in the development of the core methods, experiments, or analysis in this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Continued Related Work: Mutual Information Mechanisms

We review several classic peer prediction mechanisms that evaluate agents by estimating mutual information, and illustrate why these mechanisms fail to achieve SD-truthfulness. First, proposed by Kong and Schoenebeck [2019], we can estimate the empirical point-wise joint distribution between $\hat{X}_{a,\cdot}$ and $\hat{X}_{b,\cdot}$, using pairs of reports from two agents, $(\hat{X}_{a,j}, \hat{X}_{b,j})_{j\in[n]}$. Note that because questions are i.i.d., we can use reports from all questions to estimate the same distribution. Then, plugging the learned empirical joint distribution into an f-divergence operator returns a (biased) estimate of the f point-wise mutual information. The following is the score for the f-mutual information mechanism [Kong and Schoenebeck, 2019]:

$$S^{fMI} = D_f\left(\Pr\left(\hat{X}_{a,\cdot}, \hat{X}_{b,\cdot}\right) \| \Pr\left(\hat{X}_{a,\cdot}\right) \Pr\left(\hat{X}_{b,\cdot}\right)\right).$$

However, because the estimation is biased, the f-mutual information mechanism is only asymptotically truthful, meaning that it requires an infinite number of i.i.d. questions to satisfy Definition 3.1. When the number of tasks is small, there is a non-zero probability that truth-telling is not the best response. For example, if Alice observes "yes" on all her questions, which happens with a positive probability, then she knows that truth-telling will always result in a score of 0 regardless of Bob's responses. This is because the estimated point-wise joint distribution is the same as the product of marginal distributions. However, by misreporting "no" on some questions, she has a non-zero probability of receiving a positive score. In comparison, such deviations are discouraged under the enforced agreement (EA) mechanism because the mechanism will randomly flip the responses on behalf of the agent (see Section 6).

Several follow-up works, such as the Determinant Mutual Information (DMI) mechanism by Kong [2020] and the pairing mechanism by Schoenebeck and Yu [2020], propose different ideas to estimate the mutual information. However, they are still not SD-truthful. Consider the strategy from Example 4.1, where Alice always reports "yes." Under both DMI and the pairing mechanism, this strategy always yields a score of zero. In contrast, similar to the CA mechanism, there is a positive probability that truth-telling may result in a negative score. Thus, this constant-reporting strategy is not dominated by truth-telling under these mutual information-based mechanisms whose score can take negative values.

B Sensitivity of Prior Mechanisms Under Partition Rounding

We now map the exemplary mechanisms discussed in Section 4 into the partition rounding framework. In particular, we specify the key parameters K, S_{inf}, S_{sup}, and $m_i(e)$ for each mechanism. Consider an information structure with joint distribution J, and the marginal distribution of Alice and Bob being M^a and M^b, respectively. Furthermore, let $M_{\sigma,\sigma'} = M^a_\sigma M^b_{\sigma'}$ denote the product of marginal distributions.

Output Agreement An individual score of OA, S_i^{OA}, is 1 if both agents agree on a question i and 0 otherwise. Therefore, OA is the partition rounding reduction of itself with $K = n$, $S_{\inf} = 0$, and $S_{\sup} = 1$. By definition, when Alice exerts effort, agents' signals on the same question is sampled from the joint distribution J; while if Alice does not exert effort, she will sample a signal from the prior, meaning that agents' signals on the same question is sampled from M. The expectation of an individual score at the effort level e is thus

$$m_i^{OA}(e) = e \cdot \Pr(X_{a,i} = X_{b,i}) + (1-e) \cdot \Pr(\mu(X_{a,i}) = \mu(X_{b,i})) = e \cdot \mathrm{tr}(J) + (1-e) \cdot \mathrm{tr}(M).$$

This means that the derivative of the expected score w.r.t. effort e is

$$\nabla m_i^{OA}(e) = \mathrm{tr}(J - M) = \mathrm{tr}(\Delta),$$

where Δ is the difference between the joint distribution and the product of marginal distributions, and $\mathrm{tr}(\cdot)$ represents the trace of a matrix, i.e. the sum of the diagonal entries.

Peer Truth Serum Similar to OA, each partition of PTS is composed of a single question so that $K = n$, $S_{\inf} = 0$, and $S_{\sup} = \frac{1}{M^a_{\sigma_{\min}}}$ where $\sigma_{\min}$ is the signal occurs with the smallest probability in prior. The expected individual score is

$$m_i^{PTS}(e) = \sum_{\sigma \in \Sigma} e \cdot \frac{J_{\sigma,\sigma}}{M^a_\sigma} + (1-e) \cdot M^b_\sigma = 1 + e \cdot \sum_{\sigma \in \Sigma} \frac{\Delta_{\sigma,\sigma}}{M^a_\sigma}; \qquad \nabla m_i^{PTS}(e) = \sum_{\sigma \in \Sigma} \frac{\Delta_{\sigma,\sigma}}{M^a_\sigma}.$$

Correlated Agreement The CA mechanism requires two questions to compute an individual score, meaning that $K = n/2$. By Proposition F.1 and Lemma 5.3, when the signal space is binary, scoring Alice using $S^{CA} = \sum_{i \in [K]} S_i^{CA}$ without rounding is SD-truthful. However, when $|\Sigma| > 2$, we have to use the partition rounding reduction of CA to obtain SD-truthful where $S_{\inf} = -1$ and $S_{\sup} = 1$. On the bonus question, the probability of receiving a score of 1 is given by $\sum_{\sigma,\sigma'} (e \cdot J_{\sigma,\sigma'} + (1-e) \cdot M_{\sigma,\sigma'})\, T_\Delta(\sigma, \sigma')$. On the penalty question, the probability of penalizing a score of 1 is given by $\sum_{\sigma,\sigma'} (e \cdot M_{\sigma,\sigma'} + (1-e) \cdot M_{\sigma,\sigma'})\, T_\Delta(\sigma, \sigma')$. Combining these,

$$m_i^{CA}(e) = e \cdot \sum_{\sigma,\sigma' \in \Sigma} \Delta_{\sigma,\sigma'}\, T_\Delta(\sigma, \sigma') = e \cdot \sum_{\sigma,\sigma' \in \Sigma} (\Delta_{\sigma,\sigma'})^+; \qquad \nabla m_i^{CA}(e) = \sum_{\sigma,\sigma' \in \Sigma} (\Delta_{\sigma,\sigma'})^+,$$

where $(x)^+ = x$ if $x > 0$ and 0 otherwise. Putting these parameters into Equation (1) returns the sensitivity of the partition rounding reduction for CA. However, note that by Proposition F.1, we do not need to round the individual CA scores when $|\Sigma| = 2$. This motivates a variant of an SD-truthful implementation of CA that may have a slightly larger sensitivity in the binary-signal setting. We will introduce and test this implementation in Section 7.

We summarize the limitations of the above three SD-truthful mechanisms.

- OA does not require the partition rounding reduction to achieve SD-truthfulness and thus is likely to have a high sensitivity. However, OA is SD-truthful only when signals are self-dominating.
- PTS requires rounding to ensure SD-truthfulness. The normalizing factor $\frac{1}{M^a_{\sigma_{\min}}}$ implies that PTS will be less sensitive when the prior distribution of signals is biased.
- For the CA mechanism, only bonus questions contribute to the sensitivity of the mechanism, while penalty questions are used to enforce truthfulness. However, CA requires two questions to compute a single individual score, which means $K = n/2$.[3] Therefore, even without rounding, e.g. when $|\Sigma| = 2$, the number of effective questions is only one half of the available questions. Moreover, the rounding process divides each score by a factor of 2, which further harms the sensitivity of the mechanism.

C The Enforced Agreement Mechanism (Continued)

C.1 Arbitrary Signal Space

Unfortunately, we show that the above idea does not generalize beyond the binary-signal setting. For a general setting with $|\Sigma| = c$, the enforced agreement mechanism can be characterized by an

[3]We emphasize that increasing K by rerunning partitions is not SD-truthful (Example 4.1).

enforced marginal distribution $\Phi = (n_0, \ldots, n_{c-1})$ and an *enforcement rule* that ensures an agent's reports to follow Φ.

Let Alice's report vector $\hat{X}_a$ have an empirical marginal distribution $\Phi_{\hat{X}_a} = (\hat{m}_0, \ldots, \hat{m}_{c-1})$. The enforcement rule can be characterized as a $c \times c$ matrix $\rho\left(\hat{X}_a\right)$ (or simply ρ if there is no confusion), such that $\rho_{i,j}$ denotes the number of randomly selected questions on which Alice reports i and is flipped to j by the mechanism. By construction, $\rho_{i,j} \in \mathbb{Z}_{\geq 0}$ for any $i, j \in \Sigma$. Furthermore, $\sum_i \rho_{i,j} = n_j$ for any $j \in \Sigma$ and $\sum_j \rho_{i,j} = \hat{m}_i$ for any $i \in \Sigma$.

Suppose Alice's signal vector X_a has a marginal distribution $\Phi_a = (m_0, \ldots, m_{c-1})$. Under the enforcement rule, while reasoning the distribution the of EA score, every (randomized) strategy can be characterized by (a mixture of) *manipulation matrix*. The entry $\vartheta_{i,j}$ of a manipulation matrix denotes the number of randomly selected questions where Alice observes i but is flipped to j either by the mechanism or by herself. Similarly, $\vartheta_{i,j} \in \mathbb{Z}_{\geq 0}$ for any $i, j \in \Sigma$, $\sum_i \vartheta_{i,j} = n_j$ for any $j \in \Sigma$, and $\sum_j \vartheta_{i,j} = m_i$ for any $i \in \Sigma$. In particular, let $\vartheta^\tau(\rho)$ denote the manipulation matrix corresponding to truth-telling under the enforcement rule ρ.

Given a manipulation matrix ϑ, the final score of Alice is thus the average of c^2 binomials:

$$S^{EA}(\theta) \sim \frac{1}{n} \sum_{i,j \in \Sigma} \mathrm{Bin}(\vartheta_{i,j}, p_{ij}),$$

where $p_{ij} = \Pr(X_b = j \mid X_a = i)$ is the conditioned probability of agents' signals.

To obtain SD-truthfulness, we have to construct an enforcement rule ρ such that for any p (under some assumptions), any X_a, and any feasible ϑ, the random variable $S^{EA}(\theta)$ is first-order stochastically dominated by $S^{EA}(\tau)$. This is particularly challenging when c is large because ϑ has $c^2 - 2c + 1$ degrees of freedom, yet we must identify a ρ such that $\vartheta^\tau(\rho)$ dominates any other ϑ.

Three signals We show that even for the three-signal setting, EA is SD-truthful only under a very special type of information structure.

Definition C.1. Agents' signals are uniformly self-predicting if $\Pr(X_b = i \mid X_a = i) > \Pr(X_b = i \mid X_a = j) = \Pr(X_b = i \mid X_a = k)$ for any $j \neq i$ and $k \neq i$.

Consider the following enforcement rule:

- *Alice over-reports one signal.* If $\hat{m}_i \geq n_i$ while $\hat{m}_j \leq n_j$ for any $j \neq i$, the enforcement rule will randomly select $n_j - \hat{m}_j$ questions on which Alice reports i and flip them to j for $j \in \Sigma$ and $j \neq i$.
- *Alice over-reports two signals.* If $\hat{m}_j < n_j$ while $\hat{m}_i \geq n_i$ for any $i \neq j$, the enforcement rule will randomly select $\hat{m}_i - n_i$ questions on which Alice reports i and flip them to j for $i \in \Sigma$ and $i \neq j$.

This enforcement rule flips the minimum number of reports to enforce Φ and thus is named the *minimal enforcement rule*, denoted as ρ^*.

Proposition C.2. *If $|\Sigma| = 3$ and signals are self-predicting, the enforcement agreement mechanism is SD-truthful for any Φ if and only if it applies the minimal enforcement rule and signals are uniformly self-predicting.*

We defer the proof to Appendix F.5. Our results imply that achieving SD-truthfulness using the EA mechanism is infeasible for arbitrary information structures beyond the binary-signal setting. Investigating alternative approaches, such as mechanism designs that leverage additional structure in the information space or stronger assumptions on agent strategies, remains an interesting direction for future work.

C.2 Truthfulness of EA

Although the EA mechanism is not SD-truthful when $|\Sigma| > 2$, we show that it is still possible to obtain truthfulness as long as the information structure is known.

Given a report vector of Alice $\hat{X}_a$ with an empirical distribution $\Phi_{\hat{X}_a} = (\hat{m}_0, \ldots, \hat{m}_{c-1})$, EA will apply the *IP enforcement rule* by solving the following integer programming.

$$\begin{aligned} \max_{\vartheta} \quad & \sum_{i,j\in\Sigma} \vartheta_{i,j}\, p_{ij} \\ \text{s.t.} \quad & \sum_{i\in\Sigma} \vartheta_{i,j} = n_j, \quad \forall j \in \Sigma, \\ & \sum_{j\in\Sigma} \vartheta_{i,j} = \hat{m}_i, \quad \forall i \in \Sigma, \\ & \vartheta_{i,j} \in \mathbb{Z}_{\geq 0}, \quad \forall i, j \in \Sigma. \end{aligned}$$

Suppose the optimal solution is $\vartheta^*(\hat{X}_a)$. The mechanism will then randomly select $\vartheta^*(\hat{X}_a)_{i,j}$ questions on which $\hat{X}_a = i$ and flip them to j, for any $i, j \in \Sigma$. After the enforcement step, Alice will be scored based on the output agreement mechanism.

Proposition C.3. *For any enforcement Φ and information structure, if $\Pr(X_b \mid X_a)$ is known, the enforcement agreement mechanism with the IP enforcement rule is truthful.*

The proof is straightforward. Note that to prove truthfulness, it is sufficient to show that truth-telling maximizes the expected score which is $\sum_{i,j\in\Sigma} \vartheta^T_{i,j}\, p_{ij}$. Because the mechanism will manipulate agents' reports in an optimal way by solving the above integer programming, no strategy can achieve a higher expected score than truth-telling.

Note that when $|\Sigma| = 2$ and signals are self-predicting, the optimal solution to the above IP does not depend on the information structure p_{ij}. In particular, the IP enforcement rule always flips the over-reported signal to enforce n_0 zeros and $n - n_0$ ones and does not flip any of the under-reported signal. This echos Theorem 6.1 which suggests the EA mechanism is SD-truthful in the binary-signal setting without the knowledge of p.

Remark C.4 (Connection to Peer Truth Serum Faltings et al. [2017]). As illustrated in Section 4.2, PTS can be viewed as an output agreement mechanism (Section 4.1) where the score for agreement is weighted inversely by each signal's prior. This re-weighting shifts the probability of agreement from the joint distribution $\Pr(X_a, X_b)$, which demands self-dominating signals to ensure truthful reporting, to the conditional distribution $\Pr(X_b \mid X_a)$, which requires self-predicting signals.

In contrast, EA avoids using the prior $\Pr(X_a)$ by controlling the marginal distribution of Alice's reports. This approach similarly transforms the probability of agreement from the joint to the conditional distribution. However, the enforcement process creates additional incentive issues when the signal space is large which requires complete knowledge of the underlying information structure. Consequently, if $|\Sigma| = 2$ and signals are self-predicting, EA can obtain a stronger truthful guarantee (SD-truthfulness) without any prior knowledge; but for $|\Sigma| > 2$, it requires more extensive structural information to guarantee truthfulness.

C.3 The Optimal Enforcement

We have established that when signals are binary and self-predicting, EA with any enforcement Φ is SD-truthful. In this subsection, we examine how the choice of enforcement, $\Phi = (n_0, n - n_0)$, influences the sensitivity of EA, and in particular, what is the sensitivity-maximizing enforcement.

Fixing an information structure, the expected score given by EA, $S^{EA}(n_0, e)$, can be viewed as a function of the number of questions where the answers are enforced to zero, n_0, and Alice's effort level, e. We first present an intermediate result suggesting that the standard deviation of S^{EA} is independent of n_0. This implies that the enforcement affects the sensitivity of EA only through $\nabla\mathbb{E}[S^{EA}(e)]$.

Lemma C.5. *For a fixed information structure, std $\left(S^{EA}(n_0, e)\right)$ is independent of n_0, i.e. std $\left(S^{EA}(n_0, e)\right)$ = std $\left(S^{EA}(n_0', e)\right)$ for any $n_0, n_0' \in \{0, 1, \ldots, n\}$.*

We next analyze how the derivative of the expected score depends on the choice of enforcement. We find that the sensitivity-maximizing enforcement n_0 is determined by a quantile of a binomial distribution with success probability $\Pr(X_a = 0)$, where the quantile depends on the information structure.

Proposition C.6. *If $|\Sigma| = 2$ and signals are self-predicting, the sensitivity-maximizing enforcement of the EA mechanism $\Phi = (n_0, n - n_0)$ satisfies that*

$$n_0 = \max\left\{k \in \{0, 1, \ldots, n\} : F(k) < \frac{\eta'_{00} - \eta'_{01}}{\eta'_{00} - \eta'_{01} + \eta'_{11} - \eta'_{10}}\right\},$$

where $F(k) = \sum_{i=0}^{k} \binom{n}{i} \Pr(X_a = 0)^i (1 - \Pr(X_a = 0))^{n-i}$ is the c.d.f. of a binomial random variable with success probability $\Pr(X_a = 0)$, and $\eta'_{ij} = \Pr(X_b = j \mid X_a = i) - \Pr(X_b = j)$.

We defer the proofs of Lemma C.5 and Proposition C.6 to Appendix F.6.

D The Matching Agreement Mechanism

The matching agreement mechanism is a follow-up of the correlated agreement mechanism aiming to address more complex cheating strategies Zhang and Schoenebeck [2023b]. Similar to CA, which determines "agreement" using the delta matrix (see Section 4.3), MA determines "agreement" using a different method depending on the information structure.

Definition D.1. Let Γ be a $|\Sigma| \times |\Sigma| \times |\Sigma|$ tensor where

$$\Gamma_{\sigma_1,\sigma_2,\sigma_3} = \Pr(X_a = \sigma_1 \mid X_b = \sigma_2) - \Pr(X_a = \sigma_1 \mid X_b = \sigma_3).$$

The agreement function is thus $T_\Gamma = \text{Sign}(\Gamma)$, i.e. $T_\Gamma(\sigma_1, \sigma_2, \sigma_3) = 1$ if $\Gamma_{\sigma_1,\sigma_2,\sigma_3} > 0$, and 0 otherwise.

Next, the mechanism randomly samples a bonus question j and a penalty question k uniformly at random. An individual score computed using these two questions is thus $S_i^{MA} = T_\Gamma(\hat{X}_{a,j}, \hat{X}_{b,j}, \hat{X}_{b,k})$. In words, the matching agreement mechanism awards Alice a score of 1 if, when predicting her response on question b, the posterior conditioned on Bob's response to the same question provides a better prediction than the posterior conditioned on Bob's response on a different question q. Similar to CA, each bonus question can be paired with $n - 1$ penalty questions. Therefore, the final score of MA is the average of $n \cdot (n - 1)$ individual scores.

The Sensitivity of MA Under the Partition Rounding Reduction We further investigate the sensitivity of the MA mechanism under the partition reduction. The MA mechanism requires two questions to compute an individual score, implying that $K = n/2$. Clearly, $S_{\inf} = 0$ and $S_{\sup} = 1$ by definition. By Equation (1), to compute the sensitivity, we only need to compute the expected individual score $m_i(e)$ and its derivative. When Alice exerts full effort, the probability of obtaining a score of 1 can be computed as follows:

$$\begin{aligned}
m_i^{MA}(1) &= \sum_{\sigma_1,\sigma_2,\sigma_3} \Pr(X_{a,j} = \sigma_1, X_{b,j} = \sigma_2, X_{b,k} = \sigma_3) \cdot T_\Gamma(\sigma_1, \sigma_2, \sigma_3) \\
&= \sum_{\sigma_1,\sigma_2,\sigma_3} J_{\sigma_1,\sigma_2} M^b_{\sigma_3} \cdot T_\Gamma(\sigma_1, \sigma_2, \sigma_3) \\
&= \frac{1}{2} \sum_{\sigma_1,\sigma_2,\sigma_3} J_{\sigma_1,\sigma_2} M^b_{\sigma_3} \cdot T_\Gamma(\sigma_1, \sigma_2, \sigma_3) + \frac{1}{2} \sum_{\sigma_1,\sigma_2,\sigma_3} J_{\sigma_1,\sigma_3} M^b_{\sigma_2} \cdot T_\Gamma(\sigma_1, \sigma_3, \sigma_2) \\
&= \frac{1}{2} + \frac{1}{2} \sum_{\sigma_1,\sigma_2,\sigma_3} \left(J_{\sigma_1,\sigma_2} M^b_{\sigma_3} - J_{\sigma_1,\sigma_3} M^b_{\sigma_2}\right) \cdot T_\Gamma(\sigma_1, \sigma_2, \sigma_3) \\
&\qquad\qquad \text{(Note that } T_\Gamma(\sigma_1, \sigma_2, \sigma_3) = 1 - T_\Gamma(\sigma_1, \sigma_3, \sigma_2).\text{)} \\
&= \frac{1}{2} + \frac{1}{2} \sum_{\sigma_1,\sigma_2,\sigma_3} M^b_{\sigma_2} M^b_{\sigma_3} \, \Gamma_{\sigma_1,\sigma_2,\sigma_3} \, T_\Gamma(\sigma_1, \sigma_2, \sigma_3) \\
&= \frac{1}{2} + \frac{1}{2} \sum_{\sigma_1,\sigma_2,\sigma_3} M^b_{\sigma_2} M^b_{\sigma_3} \, (\Gamma_{\sigma_1,\sigma_2,\sigma_3})^+
\end{aligned}$$

where $(x)^+ = x$ if $x > 0$ and 0 otherwise.

When Alice exerts no effort,

$$\begin{aligned}
m_i^{MA}(0) &= \sum_{\sigma_1,\sigma_2,\sigma_3} \Pr(\hat{X}_{a,j}=\sigma_1, X_{b,j}=\sigma_2, X_{b,k}=\sigma_3)\cdot T_\Gamma(\sigma_1,\sigma_2,\sigma_3) \\
&= \sum_{\sigma_1,\sigma_2,\sigma_3} M^a_{\sigma_1} M^b_{\sigma_2} M^b_{\sigma_3}\cdot T_\Gamma(\sigma_1,\sigma_2,\sigma_3) \\
&= \frac{1}{2}\sum_{\sigma_1,\sigma_2,\sigma_3} M^a_{\sigma_1} M^b_{\sigma_2} M^b_{\sigma_3}\cdot T_\Gamma(\sigma_1,\sigma_2,\sigma_3) + \frac{1}{2}\sum_{\sigma_1,\sigma_2,\sigma_3} M^a_{\sigma_1} M^b_{\sigma_2} M^b_{\sigma_3}\cdot T_\Gamma(\sigma_1,\sigma_3,\sigma_2) \\
&= \frac{1}{2} + \frac{1}{2}\sum_{\sigma_1,\sigma_2,\sigma_3} \left(M^a_{\sigma_1} M^b_{\sigma_2} M^b_{\sigma_3} - M^a_{\sigma_1} M^b_{\sigma_2} M^b_{\sigma_3}\right)\cdot T_\Gamma(\sigma_1,\sigma_2,\sigma_3) \\
&\qquad\qquad \text{(Note that } T_\Gamma(\sigma_1,\sigma_2,\sigma_3) = 1 - T_\Gamma(\sigma_1,\sigma_3,\sigma_2)\text{.)} \\
&= \frac{1}{2}
\end{aligned}$$

The expected score while exerting effort e is a convex combination of $m_i^{MA}(1)$ and $m_i^{MA}(0)$, i.e. $m_i^{MA}(e) = \frac{1}{2} + \frac{e}{2}\sum_{\sigma_1,\sigma_2,\sigma_3} M^b_{\sigma_2} M^b_{\sigma_3} \ (\Gamma_{\sigma_1,\sigma_2,\sigma_3})^+$, and $\nabla m_i^{MA}(e) = \frac{1}{2}\sum_{\sigma_1,\sigma_2,\sigma_3} M^b_{\sigma_2} M^b_{\sigma_3} \ (\Gamma_{\sigma_1,\sigma_2,\sigma_3})^+$.

When signals are binary, we find that MA reduces to CA.

Lemma D.2. *When signals are binary,* $\sum_{\sigma_1,\sigma_2,\sigma_3} M^b_{\sigma_2} M^b_{\sigma_3} \ (\Gamma_{\sigma_1,\sigma_2,\sigma_3})^+ = \sum_{\sigma_1,\sigma_2} (\Delta_{\sigma_1,\sigma_2})^+$.

Proof. First note that by definition, $\Gamma_{\sigma_1,\sigma_2,\sigma_3} = 0$ when $\sigma_2 = \sigma_3$. Therefore, when signals are binary,

$$\begin{aligned}
\sum_{\sigma_1,\sigma_2,\sigma_3} M^b_{\sigma_2} M^b_{\sigma_3} \ (\Gamma_{\sigma_1,\sigma_2,\sigma_3})^+ &= \sum_{\sigma_1,\sigma_2} M^b_{\sigma_2} M^b_{1-\sigma_2} \ (\Gamma_{\sigma_1,\sigma_2,1-\sigma_2})^+ \\
&= \sum_{\sigma_1,\sigma_2,\sigma_3} M^b_{\sigma_2} M^b_{\sigma_3} \ (\Gamma_{\sigma_1,\sigma_2,\sigma_3})^+ \\
&= \sum_{\sigma_1,\sigma_2} (J_{\sigma_1,\sigma_2} M^b_{1-\sigma_2} - J_{\sigma_1,1-\sigma_2} M^b_{\sigma_2})^+ \\
&= \sum_{\sigma_1,\sigma_2} (J_{\sigma_1,\sigma_2} \ (1 - M^b_{\sigma_2}) - ((M^a_{\sigma_1} - J_{\sigma_1,\sigma_2}) \ M^b_{\sigma_2})^+ \\
&\qquad\qquad \text{(Signals are binary.)} \\
&= \sum_{\sigma_1,\sigma_2} (J_{\sigma_1,\sigma_2} - M^a_{\sigma_1} M^b_{\sigma_2})^+ \\
&= \sum_{\sigma_1,\sigma_2} (\Delta_{\sigma_1,\sigma_2})^+
\end{aligned}$$

□

An immediate consequence of Lemma D.2 is that the expected score under MA is a linear transformation of the expected score under CA. This further implies that the partition rounding reductions of MA and CA has the same sensitivity under the binary-signal setting.

Proposition D.3. *When signals are binary,* $\delta^{MA}(e) = \delta^{CA}(e)$ *under the partition rounding reduction.*

Proof. By Lemma D.2, we know that

$$m_i^{MA}(e) = \frac{1}{2} + \frac{e}{2}\sum_{\sigma_1,\sigma_2,\sigma_3} M^b_{\sigma_2} M^b_{\sigma_3} \ (\Gamma_{\sigma_1,\sigma_2,\sigma_3})^+ = \frac{1}{2} + \frac{e}{2}\sum_{\sigma_1,\sigma_2} (\Delta_{\sigma_1,\sigma_2})^+ = \frac{1}{2} + \frac{1}{2} m_i^{CA}(e),$$

$$\nabla m_i^{MA}(e) = \frac{1}{2}\nabla m_i^{CA}(e).$$

By Equation (1),

$$\begin{aligned}
\delta^{MA}(e) &= \frac{\nabla m_i^{MA}(e) \cdot \sqrt{n/2}}{\sqrt{\left(m_i^{MA}(e) - S_{\inf}^{MA}\right)\left(S_{\sup}^{MA} - m_i^{MA}(e)\right)}} \\
&= \frac{\frac{1}{2}\nabla m_i^{CA}(e) \cdot \sqrt{n/2}}{\sqrt{\left(\frac{1}{2}m_i^{MA}(e) + \frac{1}{2}\right)\left(\frac{1}{2} - \frac{1}{2}m_i^{MA}(e)\right)}} \\
&= \frac{\nabla m_i^{CA}(e) \cdot \sqrt{n/2}}{\sqrt{\left(m_i^{MA}(e) + 1\right)\left(1 - m_i^{CA}(e)\right)}} \\
&= \delta^{CA}(e).
\end{aligned}$$

□

E Experiments (Continue)

All codes for our experiments are provided in `https://github.com/DavidXu999/Stochastically-Dominant-Peer-Prediction`.

E.1 Information Structure

For Section 7.2 and E.3, we consider two real-world crowdsourcing datasets, each having a signal space of size $|\Sigma_1| = 2$ and $|\Sigma_2| = 4$. The information structure between a pair of agents can be characterized by a Dawid-Skene model Dawid and Skene [1979]. In particular, each task has an underlying ground truth ω that is i.i.d. sampled from a prior distribution W. Conditioning on ω, each agent receives a signal σ according to conditioned distribution Γ, where $\Gamma_{i,j} = \Pr[\sigma = j \mid \omega = i]$.

In the first dataset [Baba et al., 2018], 18 agents each labels whether a compound is appropriate or inappropriate to be synthesized. The prior of the ground truth and the conditioned signal distribution are

$$W_1 = \begin{bmatrix} 0.613 & 0.387 \end{bmatrix}, \qquad \Gamma_1 = \begin{bmatrix} 0.905 & 0.095 \\ 0.283 & 0.717 \end{bmatrix}.$$

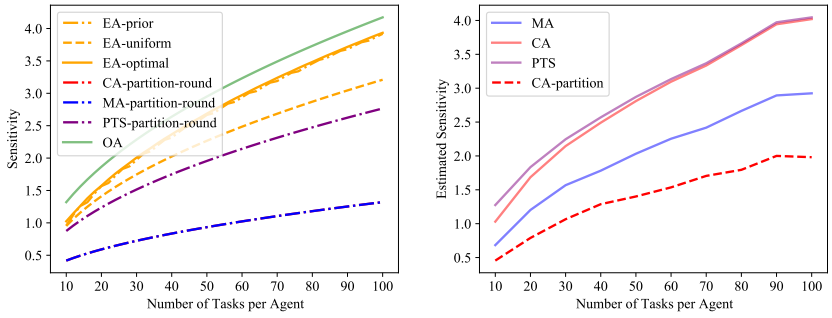
In the second dataset Venanzi et al. [2015], 110 agents provide the annotations of the sentiment of 300 tweets. There are four ground truth labels and four possible signals where the information structure is

$$w_2 = \begin{bmatrix} 0.196 & 0.241 & 0.247 & 0.316 \end{bmatrix}, \qquad \Gamma_2 = \begin{bmatrix} 0.770 & 0.122 & 0.084 & 0.024 \\ 0.091 & 0.735 & 0.130 & 0.044 \\ 0.033 & 0.062 & 0.866 & 0.039 \\ 0.068 & 0.164 & 0.099 & 0.669 \end{bmatrix}.$$

E.2 The Sensitivity of Original Implementations

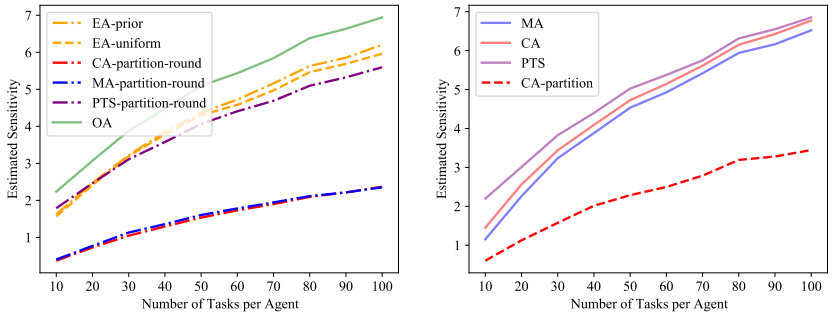
How much does the partition rounding reduction harm the sensitivity? We compare the partition rounding reduction implementations of CA, MA, and PTS with their original implementations in Figure 2. To be clear, other than the implementations considered in Section 7, we consider the following implementation of the mechanisms.

- **PTS** corresponds to the peer truth serum mechanism introduced in Section 4.2, which is truthful but not SD-truthful.
- **CA** corresponds to the original implementation of the Correlated Agreement mechanism described in Section 4.3, which is truthful but not SD-truthful.
- **CA-partition** represents the implementation where we get $K = n/2$ partitions of questions and score agents the average of the individual CA score for each partition (without normalization). By Proposition F.1, CA-partition is SD-truthful when signals are binary.
- **MA** corresponds to the original implementation of the Matching Agreement mechanism described in Appendix D, which is truthful but not SD-truthful.



(a) Mechanisms shown in Section 7 under J_1

(b) Other Mechanisms under J_1



(c) Mechanisms shown in Section 7 under J_2

(d) Other Mechanisms under J_2

Figure 2: A Comparison of the Sensitivity of Different Mechanisms.

We observe that, relative to the original implementations (solid lines), the partition-rounding reductions (dashed lines) consistently exhibit lower sensitivity. This observation further demonstrates the superior sensitivity of EA mechanisms, as it has the property of SD-truthfulness in the binary setting with no need for partition rounding reduction.

E.3 Budgetary Efficiency

Zhang and Schoenebeck [2023a] show that applying tournaments on the peer prediction scores, though it is not truthful, can incentivize a desired effort level at a lower cost of budget compared to linear payments under their model. Now, we have some more options—several SD-truthful mechanisms—that are truthful under the non-linear tournaments. In this section, we compare the performance of SD-truthful mechanisms with tournaments and truthful mechanisms with linear payments. We consider *budgetary efficiency* as the performance metric, which is defined as the minimum cost of budget to incentivize a desired effort at the symmetric equilibrium Zhang and Schoenebeck [2023a].

Formally, we consider the following payment schemes:

- The linear payment scheme pays agent i with $\pi_i = a \cdot S_i^M + b$.
- The winner-take-all tournament pays agent i with $\pi_i = C_{\text{WTA}} \cdot \mathbb{1}[S_i^M > S_j^M,\ \forall j \neq i]$.
- The Borda-count tournament pays agent i with the number of agents she beats, i.e.,

$$\pi_i = C_{\text{Borda}} \cdot \#\text{beaten} = C_{\text{Borda}} \sum_{j \neq i} \mathbb{1}[S_i^M > S_j^M]$$

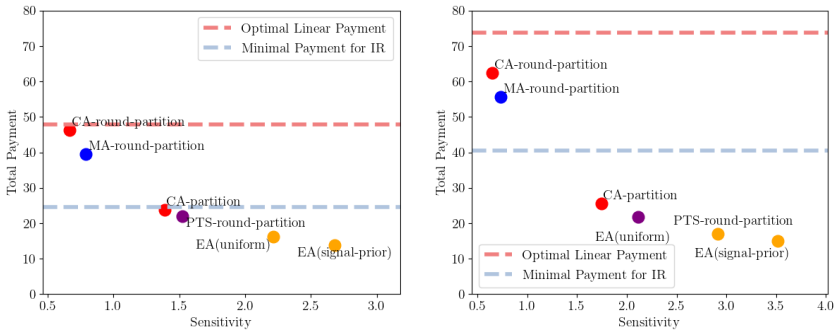
where $a, b, C_{\text{WTA}}, C_{\text{Borda}}$ are constant parameters. When there's a tie, we equally split the payoffs.

We now introduce the simulation workflow with information structure J_1. There are 50 agents and 500 tasks, where each agent works on 50 tasks and each task is randomly assigned to 5 agents. Each agent has an effort level e_i with a convex cost function $c(e_i) = e_i^2$ reflecting the increasing

marginal cost of effort. We require limited liability (i.e. the ex-post payment must be non-negative) and individual rationality (IR, i.e. the expected utility must be non-negative).

We aim to incentivize a desired effort level e as a local symmetric equilibrium where no agent wants to deviate from e. This allows us to compute the parameters in payment functions such that $\nabla\pi_i = \nabla c(e_i)\forall i$. To compute $\nabla\pi_i$, we apply the same workflow in the above subsections. In particular, we simulate samples of signals and apply various mechanisms to compute the peer prediction score $S_i^{\mathcal{M}}$ for each agent. Then, varying the effort from e to $e - \Delta e$, we can estimate ∇S_i^M and $\nabla\pi_i$ using a sample size $T = 10{,}000$.

As we mainly focus on whether the SD-Truthful mechanisms + tournament payment can achieve better budgetary efficiency than truthful mechanisms + linear payment, we compute the optimal linear payment among all truthful mechanisms mentioned in Section 7.1, and compare that with the tournament payments with all SD-Truthful mechanisms. We present the results for the winner-take-all tournament in Figure 3 and defer the figures for the Borda-count tournament to the appendix.

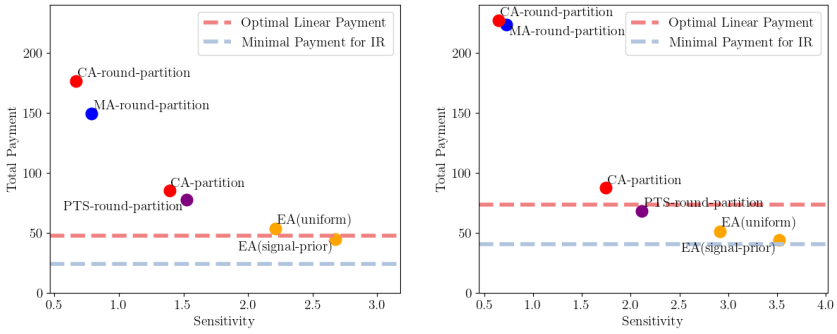


(a) Effort $e = 0.7$ (b) Effort $e = 0.9$

Figure 3: Estimated Sensitivity v.s. Total Payment of Winner-take-all Tournament

We have the following observations: 1) sensitivity aligns well with the total payment, 2) EA consistently performs the best, 3) when eliciting a high desired effort, tournament payments provide greater budget efficiency compared to linear payments, and 4) when using an efficient tournament payment, e.g. winner-take-all, the partition rounding reduction outperforms the original mechanism paired with a linear payment scheme.

E.4 Figures for Borda-count Tournaments



(a) Effort $e = 0.7$ (b) Effort $e = 0.9$

Figure 4: Estimated Sensitivity v.s. Total Payment of Borda-count Tournament

F Additional Proofs

F.1 The CA Mechanism

We first show that a single draw of the CA score is SD-truthful when $|\Sigma| = 2$.

Proposition F.1. *When $|\Sigma| = 2$, instead of taking the average, the mechanism that scores agents using an individual CA score S_i^{CA} is SD-truthful.*[4]

Proof of Proposition F.1. We first show that the agreement function T_Δ is an identity matrix if and only if signals are self-predicting. For simplicity, let $J_{\sigma,\sigma'} = \Pr(X_a = \sigma, X_b = \sigma')$ be the joint distribution, and let $M_\sigma^a = \Pr(X_a = \sigma)$ and $M_\sigma^b = \Pr(X_b = \sigma)$ be the marginal distribution of Alice and Bob respectively. The delta matrix can be written as

$$\begin{aligned}
\Delta_{\sigma,\sigma} &= J_{\sigma,\sigma} - M_\sigma^a\ M_\sigma^b \\
&= J_{\sigma,\sigma} - M_\sigma^a\ (J_{\sigma,\sigma} + J_{\sigma',\sigma}) \\
&= J_{\sigma,\sigma}\ M_{\sigma'}^a - J_{\sigma',\sigma}\ M_\sigma^a \\
&= M_\sigma^a\ M_{\sigma'}^a\ (\Pr(X_b = \sigma \mid X_a = \sigma) - \Pr(X_b = \sigma \mid X_a = \sigma'))\,. \\
\Delta_{\sigma,\sigma'} &= J_{\sigma,\sigma'} - M_\sigma^a\ M_{\sigma'}^b \\
&= J_{\sigma,\sigma'} - M_\sigma^a\ (J_{\sigma,\sigma'} + J_{\sigma',\sigma'}) \\
&= J_{\sigma,\sigma'}\ M_{\sigma'}^a - J_{\sigma',\sigma'}\ M_\sigma^a \\
&= M_\sigma^a\ M_{\sigma'}^a\ (\Pr(X_b = \sigma' \mid X_a = \sigma) - \Pr(X_b = \sigma' \mid X_a = \sigma'))\,.
\end{aligned}$$

Next, we reason about the distribution of scores under different strategies. Note that a single draw of the CA score can be -1, 0, or 1. Therefore, to prove SD-truthfulness for a single draw of the CA score, it is sufficient to show that $\Pr\left(S_i^{CA}(\tau) = 1\right) \geq \Pr\left(S_i^{CA}(\theta) = 1\right)$ and $\Pr\left(S_i^{CA}(\tau) = -1\right) \leq \Pr\left(S_i^{CA}(\theta) = -1\right)$ for any strategy θ. When signals are binary, there are only three types of untruthful strategy: 1) always reporting 0, 2) always reporting 1, and 3) always flipping 0 to 1 and 1 to 0. Any other strategy is a mixture of truth-telling and these strategies. Now, we show that all of these strategies are stochastically dominated by truth-telling by discussing two cases.

When signals are self-predicting, i.e. $\Pr(X_b = \sigma \mid X_a = \sigma) - \Pr(X_b = \sigma \mid X_a = \sigma') > 0$, the distribution of S_i^{CA} is

$$\begin{aligned}
\Pr\left(S_i^{CA}(\tau) = 1\right) &= J_{0,0}\ M_1^b + J_{1,1}\ M_0^b, \\
\Pr\left(S_i^{CA}(\tau) = -1\right) &= J_{0,1}\ M_0^b + J_{1,0}\ M_1^b.
\end{aligned}$$

Because Bob's signal is self-predicting, $\Pr\left(S_i^{CA}(\tau) = 1\right) > \Pr\left(S_i^{CA}(\tau) = -1\right)$. Let θ_f be the strategy where Alice always flips the signals.

$$\begin{aligned}
\Pr\left(S_i^{CA}(\theta_f) = 1\right) &= J_{0,1}\ M_0^b + J_{1,0}\ M_1^b = \Pr\left(S_i^{CA}(\tau) = -1\right), \\
\Pr\left(S_i^{CA}(\theta_f) = -1\right) &= J_{0,0}\ M_1^b + J_{1,1}\ M_0^b = \Pr\left(S_i^{CA}(\tau) = 1\right).
\end{aligned}$$

This means that $S_i^{CA}(\theta_f)$ is dominated by $S_i^{CA}(\tau)$.

Let μ be the strategy where Alice always reports the same signal.

$$\Pr\left(S_i^{CA}(\mu) = 1\right) = \Pr\left(S_i^{CA}(\mu) = -1\right) = M_0^b\ M_1^b.$$

Therefore,

$$\begin{aligned}
\Pr\left(S_i^{CA}(\tau) = 1\right) - \Pr\left(S_i^{CA}(\mu) = 1\right) &= J_{0,0}\ M_1^b + J_{1,1}\ M_0^b - M_0^b\ M_1^b \\
&= J_{0,0}\ M_1^b + J_{1,1}\ M_0^b - M_0^b\ (J_{0,1} + J_{1,1}) \\
&= J_{0,0}\ M_1^b - J_{0,1}\ M_0^b \\
&= M_0^b\ M_1^b\ (\Pr(X_a = 0 \mid X_b = 0) - \Pr(X_b = 0 \mid X_a = 1) \\
&> 0
\end{aligned}$$

[4] When $|\Sigma| \geq 3$, counter-examples exist showing that scoring agents with S_i^{CA} is no longer SD-truthful for general information structure.

Similarly, we can observe that $\Pr\left(S_i^{CA}(\tau) = -1\right) < \Pr\left(S_i^{CA}(\mu) = -1\right)$.

The analysis for the case when signals are not self-predicting, i.e. $T_\Delta(\sigma, \sigma') = 1$ if and only if $\sigma \neq \sigma'$, is analogous. □

Next, we show that when $|\Sigma| = 3$, counterexamples exist showing that truth-telling does not FOSD every untruthful strategy.

Example F.1. Consider the following joint distribution:

$$J = \begin{pmatrix} 0.02 & 0.01 & 0.02 \\ 0.2 & 0.4 & 0.25 \\ 0.00 & 0.05 & 0.05 \end{pmatrix}.$$

The marginal distribution of Alice and Bob are

$$M^a = (0.05, \quad 0.85, \quad 0.1) \text{ and } M^b = (0.22, \quad 0.46, \quad 0.32).$$

The product of marginal distribution and the corresponding agreement function are

$$\Delta = \begin{pmatrix} 0.009 & -0.013 & 0.004 \\ 0.013 & 0.009 & -0.022 \\ -0.022 & 0.004 & 0.018 \end{pmatrix}, \quad T_\Delta = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

When Alice reports truthfully, the probability of obtaining a score of 1, i.e. obtaining a score of 1 on the bonus question and obtaining a score of 0 on the penalty question, is given by

$$\Pr(S_i^{CA}(\tau) = 1) = \underbrace{(J_{0,0} + J_{0,2}) \cdot M_1^b}_{\text{Alice reports 0}} + \underbrace{(J_{1,0} + J_{1,1}) \cdot M_2^b}_{\text{Alice reports 1}} + \underbrace{(J_{2,1} + J_{2,2}) \cdot M_0^b}_{\text{Alice reports 0}} = 0.2324.$$

However, when Alice plays a strategy μ that always reports 0, the probability of obtaining a score of 1 is given by

$$\Pr(S_i^{CA}(\mu) = 1) = (M_0^b + M_2^b) \cdot M_1^b = 0.2474 > \Pr(S_i^{CA}(\tau) = 1).$$

Therefore, μ is not first-order stochastically dominated by τ.

F.2 The Direct Rounding Reduction

Proof of Proposition 5.2. By definition, the sensitivity of a mechanism with score $S(e)$ at effort level e is

$$\delta^{\mathcal{M}}(e) = \frac{\nabla m^{\mathcal{M}}(e)}{\text{std}^{\mathcal{M}}(e)}.$$

The direct rounding of $S^{\mathcal{M}}(e)$, denoted as $\tilde{S}^{\mathcal{M}}(e)$, is a Bernoulli variable with success probability $\mathbb{E}[\lambda^{\mathcal{M}}] = \frac{m^{\mathcal{M}}(e) - S_{\inf}^{\mathcal{M}}}{S_{\sup}^{\mathcal{M}} - S_{\inf}^{\mathcal{M}}}$. Therefore,

$$\begin{aligned} &\nabla \mathbb{E}\left[\tilde{S}^{\mathcal{M}}(e)\right] = \frac{1}{S_{\sup}^{\mathcal{M}} - S_{\inf}^{\mathcal{M}}} \cdot \nabla m^{\mathcal{M}}(e) \\ &\text{std}\left(\tilde{S}^{\mathcal{M}}(e)\right) = \sqrt{\mathbb{E}[\lambda^{\mathcal{M}}]\left(1 - \mathbb{E}[\lambda^{\mathcal{M}}]\right)} = \frac{1}{S_{\sup}^{\mathcal{M}} - S_{\inf}^{\mathcal{M}}} \cdot \sqrt{\left(m^{\mathcal{M}}(e) - S_{\inf}^{\mathcal{M}}\right)\left(S_{\sup}^{\mathcal{M}} - m^{\mathcal{M}}(e)\right)} \\ \Rightarrow \quad &\tilde{\delta}^{\mathcal{M}}(e) = \frac{\nabla m^{\mathcal{M}}(e)}{\sqrt{\left(m^{\mathcal{M}}(e) - S_{\inf}^{\mathcal{M}}\right)\left(S_{\sup}^{\mathcal{M}} - m(e)\right)}}. \end{aligned} \tag{3}$$

□

F.3 The Partition Rounding Reduction

Proof of Lemma 5.3. Let $S_i(\tau)$ and $S_i(\theta)$ be the random variable of each individual score under truth-telling and strategy θ, respectively. $S_i(\tau)$ (and $S_i(\theta)$) is i.i.d. because questions are assumed to be i.i.d. and strategies are task-independent. We want to show

$$S_i(\tau) \quad \succeq_{\text{FOSD}} \quad S_i(\theta) \qquad \Leftrightarrow \qquad S(\tau) \quad \succeq_{\text{FOSD}} \quad S(\theta).$$

Forward Direction. Suppose $S_i(\tau) \succeq_{\text{FOSD}} S_i(\theta)$ for any θ. By definition, this means for each individual score, we can find a coupling such that whenever $S_i(\theta) = s_i$, $S_i(\tau) = s_i' \geq s_i$. Applying this coupling for every individual score and taking the average gives us a coupling for $S(\tau)$ and $S(\theta)$: whenever $S(\theta) = \sum_{i\in[K]} s_i$, $S(\tau) = \sum_{i\in[K]} s_i' \geq \sum_{i\in[K]} s_i$ in the coupling. Therefore, in this coupling, $\Pr(S(\theta) \leq t) \geq \Pr(S(\tau) \leq t)$ for any $t \in [K \cdot S_{\inf}, K \cdot S_{\sup}]$ where $S_{\inf}$ and $S_{\sup}$ are the infimun and supremum of each S_i, respectively. This completes the proof for the sufficiency.

Reverse Direction. We prove the contrapositive of the statement: if $S_i(\tau)$ does not first-order stochastic dominates $S_i(\theta)$, then $S(\tau)$ does not first-order stochastic dominates $S(\theta)$. By the failure of FOSD for the individuals, there exists some threshold t_1 and (by the i.i.d. assumption) every i such that $F_\tau(t_1) > F_\theta(t_1)$, where F_τ and F_θ denote that c.d.f. for $S_i(\tau)$ and $S_i(\theta)$ respectively. Let F_τ^k and F_θ^k be the c.d.f. of the sum of k i.i.d. individual scores, i.e. $\sum_{i\in[k]} S_i(\tau)$ and $\sum_{i\in[k]} S_i(\theta)$ respectively. We want to find a t_K such that $F_\tau^K(t_K) > F_\theta^K(t_K)$. Then, because S is a linear scaling (by $1/K$) of the sum, the inequality transfers directly to the averages.

We prove this via an inductive argument:

- Base case ($K = 1$): by assumption, there exists a t_1 such that $F_\tau(t_1) > F_\theta(t_1)$.
- Inductive step: suppose that for $K-1$ there exists some threshold t_{K-1} such that

  $$F_\tau^{K-1}(t_{K-1}) > F_\theta^{K-1}(t_{K-1}).$$

  Now, consider the K-fold convolution. Note that the convolution of discrete distributions is a weighted sum of the individual c.d.f..

  $$F_\tau^K(t_K) = \sum_s \Pr\left(\sum_{i\in[K-1]} S_i(\tau) = s\right) F_\tau(t_K - s),$$

  and similarly for $F_\theta^K(t_K)$. By the base case and the induction assumption, we can set $t_K = t_{K-1} + t_1$ so that $F_\tau^K(t_K) > F_\theta^K(t_K)$. This means that there exists a threshold $t = t_K/K$ so that

  $$\Pr\big(S(\tau) \leq t\big) > \Pr\big(S(\theta) \leq t\big).$$

This completes the proof of the necessity. □

F.4 EA in the Binary-signal Setting

Proof of Theorem 6.1. Suppose Alice observes m_0 zeros and m_1 ones on n questions while the enforced empirical distribution is $\Phi = (n_0, n_1)$. Alice's strategy and Φ thus determines the number of i.i.d. samples from each of the four binomials.

To prove SDT, we have to show that truth-telling results in a score distribution that first-order stochastic dominates the score distribution under any other strategy. Suppose W.L.O.G. that $m_0 > n_0$ and the mechanism will randomly flip Alice's reports on $m_0 - n_0$ questions from 0 to 1. If Alice reports truthfully, there are n_0 questions where Alice reports 0 and will be scored 1 if Bob also reports 0; there are m_1 questions where Alice reports 1 and will be scored 1 if Bob also reports 1; and there are $m_0 - n_0$ questions where Alice observes 0 but her reports are flipped to 1 which means that she will be scored 1 if Bob reports 1 on those questions. Therefore, the final score of Alice can be expressed as

$$n \cdot S(\tau) \sim \text{Bin}(n_0, p_{00}) + \text{Bin}(m_1, p_{11}) + \text{Bin}(m_0 - n_0, p_{01}). \tag{4}$$

To show that $S(\tau)$ dominates $S(\theta)$ for any strategy θ, it is sufficient to focus on deterministic strategies. Suppose $x_{i,j}$ is the number of questions that Alice observes i but reports j for $i, j \in \{0, 1\}$. First note that $x_{0,0} + x_{0,1} = m_0$ and $x_{1,0} + x_{1,1} = m_1$. Then, we argue that it is sufficient to focus on strategies that satisfy $x_{0,0} + x_{1,0} = n_0$ and $x_{0,1} + x_{1,1} = n_1$. Otherwise, the mechanism will enforce Φ on behalf of Alice which is identical for Alice to flip her signals by herself. Fixing any Φ and Φ_{X_a}, the above four constraints suggest that there is only one free variable in x, meaning that there is only one possible way to cheat—altering $x_{0,0} \in \{0, \dots, n_0\}$.

Now, we show that reporting $x_{0,0} = n_0$ is the best strategy—reporting any $x_{0,0}$ results in a score distribution dominated by reporting $x_{0,0} = n_0$. Also, note that truth-telling is one of the strategies that has $x_{0,0} = n_0$. Let $x_{0,0} = n_0 - k$ where $k \geq 0$. Then, the final score of Alice and be expressed as

$$n \cdot S(\theta) \sim \text{Bin}(n_0 - k, p_{00}) + \text{Bin}(m_1 - k, p_{11}) + \text{Bin}(m_0 - n_0 + k, p_{01}) + \text{Bin}(k, p_{10}). \quad (5)$$

Compared with the truth-telling score distribution in Equation (2), any untruthful reporting can be viewed as subtracting k samples from $\text{Bin}(n_0, p_{00})$ and draw k more samples from $\text{Bin}(k, p_{10})$; subtracting k samples from $\text{Bin}(m_1, p_{11})$ and draw k more samples from $\text{Bin}(m_0 - n_0 + k, p_{01})$.

When signals are self-predicting, $p_{00} > p_{10}$ and $p_{11} > p_{01}$. We can construct a coupling to show that $S(\theta)$ is dominated by $S(\tau)$. In particular, for the k questions whose scores are sampled from $\text{Bin}(k, p_{10})$, we couple them with k out of the n_0 samples from $S(\theta) \sim \text{Bin}(n_0 - k, p_{00})$; for the $n_0 - k$ questions whose scores are sampled from $\sim \text{Bin}(n_0 - k, p_{00})$, we couple them with the remaining $n_0 - k$ samples from $\sim \text{Bin}(n_0 - k, p_{00})$; we do the same for the other two Binomial distributions. Under this coupling, it is clear that making $x_{0,0}$ as large as possible will have the dominating score distribution, which happens while truth-telling. □

F.5 EA in the Three-signal Setting

Proof of Proposition C.2.

Sufficiency. We prove the sufficiency by discussing two cases. First, suppose W.L.O.G. that $m_0 \geq n_0$, $m_1 \leq n_1$, and $m_2 \leq n_2$. Under the minimal enforcement rule, the manipulation matrix corresponds to truth-telling is

$$\vartheta^\tau(\rho^*) = \begin{bmatrix} n_0 & n_1 - m_1 & n_2 - m_2 \\ 0 & m_1 & 0 \\ 0 & 0 & m_2 \end{bmatrix}.$$

When we score an arbitrary report vector $\hat{X}_a$, it is sufficient to consider its manipulation matrix ϑ. Therefore, we want to show that ϑ^τ will result in a score distribution that first-order stochastically dominates the score distribution of any other feasible ϑ. Note that there are 9 variables in a 3×3 matrix and 5 constraints: the sum of variables in row i is m_i and the sum of variables in column j is n_j while only 5 of these 6 constraints are independent. This means that any manipulation matrix be characterized by 4 independent variables denoted as $t_1, \dots t_4 \geq 0$.

$$\vartheta = \begin{bmatrix} n_0 - t_1 - t_2 & n_1 - m_1 + t_1 + t_4 - t_3 & n_2 - m_2 + t_2 + t_3 - t_4 \\ t_1 & m_1 - t_1 - t_4 & t_4 \\ t_2 & t_3 & m_2 - t_2 - t_3 \end{bmatrix}.$$

The variables $t_1, \dots t_4$ must satisfy the constraints ensuring that every entry in ϑ remains non-negative.

The final score of truth-telling is given by the average of n Bernoulli variables where $\vartheta^\tau(\rho^*)_{i,j}$ of these variables are sampled from $\text{Bernoulli}(p_{i,j})$. Similarly, the final score corresponding to the manipulation matrix ϑ is the average of n Bernoulli variables where $\vartheta^\theta_{i,j}$ of them are sampled from $\text{Bernoulli}(p_{i,j})$.

Our goal is to construct a coupling of these Bernoulli variables such that, for any given Φ_{X_a}, Φ, and $t_1, \dots, t_4$, each variable under truth-telling with manipulation matrix $\vartheta^\tau(\rho^*)$ has a weakly higher probability of success than its coupled counterpart under ϑ. We denote this coupling by $Z^\tau_{i,j} \overset{k}{\to} Z^\theta_{i',j'}$ to indicate that we pair k Bernoulli variables associated with $\vartheta^\tau(\rho^*)$—each drawn from $\text{Bernoulli}(p_{ij})$—with k Bernoulli variables associated with ϑ—each drawn from $\text{Bernoulli}(p_{i'j'})$.

Our coupling is:

$$\begin{bmatrix} Z^\tau_{0,0} \xrightarrow{n_0-t_1-t_2} Z^\theta_{0,0} & Z^\tau_{1,1} \xrightarrow{t_1+t_4} Z^\theta_{0,1} & Z^\tau_{2,2} \xrightarrow{t_2+t_3} Z^\theta_{0,2} \\ Z^\tau_{0,0} \xrightarrow{t_1} Z^\theta_{1,0} & Z^\tau_{1,1} \xrightarrow{m_1-t_1-t_4} Z^\theta_{1,1} & Z^\tau_{0,2} \xrightarrow{t_4} Z^\theta_{1,2} \\ Z^\tau_{0,0} \xrightarrow{t_2} Z^\theta_{2,0} & Z^\tau_{0,1} \xrightarrow{t_3} Z^\theta_{2,1} & Z^\tau_{2,2} \xrightarrow{m_2-t_2-t_3} Z^\theta_{2,2} \end{bmatrix}.$$

Under this coupling, every diagonal variable $Z^\theta_{i,i}$ is coupled with $Z^\tau_{i,i}$ which is drawn from the same Bernoulli distribution. Furthermore, every non diagonal variable $Z^\theta_{i,j}$ is either coupled with $Z^\tau_{i,i}$ which has a larger success probability because of self-prediction, or coupled with $Z^\tau_{k,j}$ which has the same success probability because signals are uniformly self-predicting.

The proof of the second case is analogous. suppose W.L.O.G. that $m_0 > n_0$, $m_1 > n_1$, and $m_2 < n_2$. Under the minimal enforcement rule, the manipulation matrix corresponds to truth-telling is

$$\vartheta^\tau = \begin{bmatrix} n_0 & 0 & m_0 - n_0 \\ 0 & n_1 & m_1 - n_1 \\ 0 & 0 & m_2 \end{bmatrix}. \tag{6}$$

Any feasible manipulation matrix can be written as

$$\vartheta^\theta = \begin{bmatrix} n_0 - t_1 - t_2 & t_4 & m_0 - n_0 + t_1 + t_2 - t_4 \\ t_1 & n_1 - t_3 - t_4 & m_1 - n_1 + t_3 + t_4 - t_1 \\ t_2 & t_3 & m_2 - t_2 - t_3 \end{bmatrix}. \tag{7}$$

We construct the following coupling. If $t_1 \geq t_4$, we apply the term in red; while if $t_1 < t_4$, we apply the term in blue.

$$\begin{bmatrix} Z^\tau_{0,0} \xrightarrow{n_0-t_1-t_2} Z^\theta_{0,0} & Z^\tau_{1,1} \xrightarrow{t_4} Z^\theta_{0,1} & Z^\tau_{2,2} \xrightarrow{t_2} Z^\theta_{0,2},\ Z^\tau_{1,2} \xrightarrow{t_1-t_4} Z^\theta_{0,2} \\ Z^\tau_{0,0} \xrightarrow{t_1} Z^\theta_{1,0} & Z^\tau_{1,1} \xrightarrow{n_1-t_3-t_4} Z^\theta_{1,1} & Z^\tau_{2,2} \xrightarrow{t_3} Z^\theta_{1,2},\ Z^\tau_{0,2} \xrightarrow{t_4-t_1} Z^\theta_{1,2} \\ Z^\tau_{0,0} \xrightarrow{t_2} Z^\theta_{2,0} & Z^\tau_{1,1} \xrightarrow{t_3} Z^\theta_{2,1} & Z^\tau_{2,2} \xrightarrow{m_2-t_2-t_3} Z^\theta_{2,2} \end{bmatrix}.$$

Under the same argument, we can observe that every $Z^\theta_{i,j}$ is coupled with a $Z^\tau_{i',j'}$ with a weakly larger success probability. This completes the proof of sufficiency.

Necessity. We first show that if the mechanism's enforcement rule is not minimal, it is not SD-truthful. For a non-minimal enforcement rule, there must exist a report vector $\hat{X}_a$ and an enforced histogram Φ such that if Alice reports $\hat{X}_a$ truthfully, there exists a diagonal value in the corresponding manipulation matrix that can be increased without breaking the constraints. In other words, there exists a strategy θ whose corresponding manipulation matrix is ϑ^θ and a column j such that $\vartheta^\theta_{j,j} > \vartheta^\tau_{j,j}$. Then, the strategy θ is not dominated by truth-telling even when signals are uniformly self-predicting. This is because by playing θ, Alice can move $\vartheta^\theta_{j,j} - \vartheta^\tau_{j,j}$ samples from $\text{Bernoulli}(p_{ij})$ or $\text{Bernoulli}(p_{kj})$ to $\text{Bernoulli}(p_{jj})$ where $i, k \neq j$. For self-predicting signals, p_{jj} is the largest column value.

Next, we show that if signals are not uniformly self-predicting, the enforced agreement mechanism is not SD-truthful even when the enforcement rule is minimal. We can easily find a counterexample by comparing Equation (6) with Equation (7). Suppose W.L.O.G. that $\vartheta^\tau_{0,2} < \vartheta^\tau_{1,2}$, then Alice can set $t_1 = t_2 = t_3 = 0$ and $t_4 = \max(n_1, m_0 - n_0)$ in which case she can move t_4 samples from $\text{Bernoulli}(p_{0,2})$ to $\text{Bernoulli}(p_{1,2})$, resulting a score distribution that dominates truth-telling. This completes the proof of necessity. □

F.6 The Optimal Enforcement

Proof of Lemma C.5. Suppose Alice observes 0 on m_0 out of the n questions and observes 1 on the remaining $n - m_0$ questions. There are two cases.

First, when $m_0 \leq n_0$, the final score of Alice is the average of the following three type of questions:

- m_0 questions where Alice observes 0 and receives a score of 1 if Bob also reports 0.
- $n - n_0$ questions where Alice observes 1 and receives a score of 1 if Bob also reports 1.

- $n_0 - m_0$ questions where Alice observes 1 but receives a score of 1 if Bob reports 0.

When Alice exerts full effort ($e = 1$), Bob's reports (which match his signals) follow the conditional probability distribution $p_{ij} = \Pr(X_b = j \mid X_a = i)$. In contrast, when Alice exerts no effort ($e = 0$), her signals are uninformative about Bob's signals, and thus her belief about Bob's reports follows the marginal distribution $M_i^b = \Pr(X_b = i)$. Therefore, conditioned on Alice observing 0 in m_0 questions, the score defined by EA follows the sum of three binomials:

$$n \cdot S^{EA}(n_0, e = 1 \mid m_0 \leq n_0) \sim \text{Bin}(m_0, p_{00}) + \text{Bin}(n - n_0, p_{11}) + \text{Bin}(n_0 - m_0, p_{10}),$$
$$n \cdot S^{EA}(n_0, e = 0 \mid m_0 \leq n_0) \sim \text{Bin}(m_0, M_0^b) + \text{Bin}(n - n_0, M_1^b) + \text{Bin}(n_0 - m_0, M_0^b).$$

Note that for any question, the probability of Alice observing 0 while exerting effort is the marginal distribution $\Pr(X_a = 0)$, which is identical to the probability of observing 0 while exerting no effort. This implies that under our definition of effort, the probability of Alice observing 0 on m_0 questions does not depend on the effort. Let $\rho_0 = \sum_{j \in [n]} (1 - X_{a,j})$ be the number of questions where Alice observes 0. The probability that Alice observes 0 on m_0 questions is

$$\Pr(\rho_0 = m_0) = \binom{n}{m_0} (M_0^a)^{m_0} (1 - M_0^a)^{n - m_0}.$$

Consequently, when $m_0 \leq n_0$, Alice's score, conditioned on observing 0 in $m_0 \leq n_0$ questions (which occurs with probability $\Pr(\rho_0 = m_0)$), follows the distribution:

$$\begin{aligned} n \cdot S^{EA}(n_0, e \mid m_0 \leq n_0) \sim &\text{Bin}\left(m_0, e\, p_{00} + (1 - e)\, M_0^b\right) + \text{Bin}\left(n - n_0, e\, p_{11} + (1 - e)\, M_1^b\right) \\ &+ \text{Bin}\left(n_0 - m_0, e\, p_{10} + (1 - e)\, M_0^b\right). \end{aligned}$$

In the second case, where $m_0 > n_0$, the final score of Alice is the average of the following three type of questions:

- n_0 questions where Alice observes 0 and receives a score of 1 if Bob also reports 0.
- $n - m_0$ questions where Alice observes 1 and receives a score of 1 if Bob also reports 1.
- $m_0 - n_0$ questions where Alice observes 0 but receives a score of 1 if Bob reports 1.

By the analogous arguments, we know that Alice's score, conditioned on observing 0 in $m_0 > n_0$ questions, follows the distribution:

$$\begin{aligned} n \cdot S^{EA}(n_0, e \mid m_0 > n_0) \sim &\text{Bin}\left(n_0, e\, p_{00} + (1 - e)\, M_0^b\right) + \text{Bin}\left(n - m_0, e\, p_{11} + (1 - e)\, M_1^b\right) \\ &+ \text{Bin}\left(m_0 - n_0, e\, p_{01} + (1 - e)\, M_1^b\right). \end{aligned}$$

Putting everything together,

$$S^{EA}(n_0, e) \sim \sum_{m_0 = 0}^{n_0} \Pr(\rho_0 = m_0) \cdot S^{EA}(n_0, e \mid m_0 \leq n_0) + \sum_{m_0 = n_0 + 1}^{n} \Pr(\rho_0 = m_0) \cdot S^{EA}(n_0, e \mid m_0 > n_0) \tag{8}$$

Now, we show that the standard deviation of $S^{EA}(n_0, e)$ is independent of n_0. For simplicity, let $\eta_{ij} = e\ p_{ij} + (1 - e)\ M_j^b$. Note that $\eta_{i0} + \eta_{i1} = e\ (p_{i0} + p_{i1}) + (1 - e)\ (M_0^b + M_1^b) = 1$. Furthermore, the variance of a binomial with parameters n, p is $np(1 - p)$. When $m_0 \leq n_0$, the variance of the final score is

$$\begin{aligned} \text{var}\left(n \cdot S^{EA}(n_0, e \mid m_0 \leq n_0)\right) &= m_0\ \eta_{00}\ (1 - \eta_{00}) + (n - n_0)\ \eta_{11}\ (1 - \eta_{11}) + (n_0 - m_0)\ \eta_{10}\ (1 - \eta_{10}) \\ &= m_0\ \eta_{00}\ (1 - \eta_{00}) + (n - m_0)\ \eta_{11}\ (1 - \eta_{11}). \end{aligned}$$

Similarly, when $m_0 > n_0$,

$$\begin{aligned} \text{var}\left(n \cdot S^{EA}(n_0, e \mid m_0 > n_0)\right) &= n_0\ \eta_{00}\ (1 - \eta_{00}) + (n - m_0)\ \eta_{11}\ (1 - \eta_{11}) + (m_0 - n_0)\ \eta_{01}\ (1 - \eta_{01}) \\ &= m_0\ \eta_{00}\ (1 - \eta_{00}) + (n - m_0)\ \eta_{11}\ (1 - \eta_{11}). \end{aligned}$$

Therefore, conditioned on observing 0 on m_0 questions, the standard deviation $\text{std}\left(S^{EA}(n_0, e \mid m_0)\right) = \sqrt{m_0\ \eta_{00}\ (1-\eta_{00}) + (n-m_0)\ \eta_{11}\ (1-\eta_{11})}$ is independent of n_0. Consequently, the standard deviation of the final score, averaging over m_0, is also independent of n_0. □

Proof of Proposition C.6. By Lemma C.5, the standard deviation of the EA mechanism is independent of the enforcement Φ. Therefore, the enforcement that maximizes the sensitivity is the one that maximizes the derivative of the expected score w.r.t. effort e. By Equation (8), the score given by EA can be expressed as the average of many binomial variables. Note that the expectation of a binomial variable with parameter (n, p) is $n \cdot p$. We thus have

$$n \cdot \mathbb{E}\left[S^{EA}(n_0, e \mid m_0 \le n_0)\right] = m_0\eta_{00} + (n-n_0)\ \eta_{11} + (n_0-m_0)\ \eta_{10}$$
$$n \cdot \mathbb{E}\left[S^{EA}(n_0, e \mid m_0 > n_0)\right] = n_0\eta_{00} + (n-m_0)\ \eta_{11} + (m_0-n_0)\ \eta_{01},$$

where $\eta_{ij} = e\ p_{ij} + (1-e)\ M_j^b$. The derivative of η_{ij} w.r.t. e is thus $\eta'_{ij} = p_{ij} - M_j^b$.

We aim to find the n_0 that maximizes

$$\frac{\partial \mathbb{E}\left[S^{EA}(n_0, e)\right]}{\partial e} = \sum_{m_0=0}^{n_0} \Pr\left(\rho_0 = m_0\right) \cdot \left(m_0\ \eta'_{00} + (n-n_0)\ \eta'_{11} + (n_0-m_0)\ \eta'_{10}\right)$$
$$+ \sum_{m_0=n_0+1}^{n} \Pr\left(\rho_0 = m_0\right) \cdot \left(n_0\ \eta'_{00} + (n-m_0)\ \eta'_{11} + (m_0-n_0)\ \eta'_{01}\right).$$

Let n_0 increase by 1:

$$\frac{\partial \mathbb{E}\left[S^{EA}(n_0+1, e)\right]}{\partial e} = \sum_{m_0=0}^{n_0+1} \Pr\left(\rho_0 = m_0\right) \cdot \left(m_0\ \eta'_{00} + (n-n_0-1)\ \eta'_{11} + (n_0+1-m_0)\ \eta'_{10}\right)$$
$$+ \sum_{m_0=n_0+2}^{n} \Pr\left(\rho_0 = m_0\right) \cdot \left((n_0+1)\ \eta'_{00} + (n-m_0)\ \eta'_{11} + (m_0-n_0-1)\ \eta'_{01}\right).$$

Therefore, the marginal return of increasing n_0 is given by

$$d\nabla S^{EA}(n_0) = \frac{\partial \mathbb{E}\left[S^{EA}(n_0+1, e)\right]}{\partial e} - \frac{\partial \mathbb{E}\left[S^{EA}(n_0, e)\right]}{\partial e}$$
$$= \sum_{m_0=0}^{n_0} \Pr\left(\rho_0 = m_0\right) \cdot \left(\eta'_{10} - \eta'_{11}\right) + \sum_{m_0=n_0+1}^{n} \Pr\left(\rho_0 = m_0\right) \cdot \left(\eta'_{00} - \eta'_{01}\right).$$

Note that

$$\begin{aligned}
\eta'_{10} - \eta'_{11} &= p_{10} - M_0^b - p_{11} + M_1^b \\
&= -2(p_{11} - M_1^b) && (p_{01} = 1 - p_{11}, \text{ and } M_0^b = 1 - M_1^b.) \\
&= -2(p_{11} - (M_0^a\ p_{01} + M_1^a\ p_{11})) \\
&= -2(p_{11}\ (1 - M_1^a) - p_{01}M_0^a) \\
&= -2\ M_0^a\ (p_{11} - p_{01}) \\
&< 0 && \text{(By self-prediction.)}
\end{aligned}$$

Similarly, we can show that $\eta'_{00} - \eta'_{01} > 0$.

Therefore, when $n_0 = 0$, $d\nabla S^{EA}(0) > 0$, and when $n_0 = n$, $d\nabla S^{EA}(n) < 0$. Furthermore, $d\nabla S^{EA}(n_0)$ decreases in n_0. The sensitivity-maximizing n_0 is thus the largest integer such that $d\nabla S^{EA}(n_0) > 0$. Let $F(k) = \sum_{m_0=0}^{k} \Pr\left(\rho_0 = m_0\right)$.

$$\begin{aligned}
& d\nabla S^{EA}(n_0) > 0 \\
\Leftrightarrow\quad & F(n_0)\left(\eta'_{10} - \eta'_{11}\right) + \left(1 - F(n_0)\right)\left(\eta'_{00} - \eta'_{01}\right) > 0 \\
\Leftrightarrow\quad & F(n_0) < \frac{\eta'_{00} - \eta'_{01}}{\eta'_{00} - \eta'_{01} + \eta'_{11} - \eta'_{10}}.
\end{aligned}$$

□