Surrogate Scoring Rules

12

YANG LIU, UC Santa Cruz, USA
JUNTAO WANG and YILING CHEN, Harvard University, USA

Strictly proper scoring rules (SPSR) are incentive compatible for eliciting information about random variables from strategic agents when the principal can reward agents after the realization of the random variables. They also quantify the quality of elicited information, with more accurate predictions receiving higher scores in expectation. In this article, we extend such scoring rules to settings in which a principal elicits private probabilistic beliefs but only has access to agents' reports. We name our solution *Surrogate Scoring Rules* (SSR). SSR is built on a bias correction step and an error rate estimation procedure for a reference answer defined using agents' reports. We show that, with a little information about the prior distribution of the random variables, SSR in a multi-task setting recover SPSR in expectation, as if having access to the ground truth. Therefore, a salient feature of SSR is that they quantify the quality of information despite the lack of ground truth, just as SPSR do for the setting *with* ground truth. As a by-product, SSR induce *dominant uniform strategy truthfulness* in reporting. Our method is verified both theoretically and empirically using data collected from real human forecasters.

CCS Concepts: • **Information systems → Incentive schemes**; • **Theory of computation → Quality of equilibria;**

Additional Key Words and Phrases: Strictly proper scoring rules, information elicitation without verification, peer prediction, dominant strategy incentive compatibility, information calibration

ACM Reference format:
Yang Liu, Juntao Wang, and Yiling Chen. 2023. Surrogate Scoring Rules. *ACM Trans. Econ. Comput.* 10, 3, Article 12 (February 2023), 36 pages.
https://doi.org/10.1145/3565559

1 INTRODUCTION

Accurate assessment of random variables of interest (e.g., how likely the S&P 500 index will go up next week) plays a crucial role in a wide array of applications, including computational finance [7], geopolitical forecasting [10, 43], weather and climate forecasting [12], and the prediction of the replicability of social science studies [1, 15]. Since such assessments are often elicited from people,

Y. Liu and J. Wang contributed equally to this research.

This research is based upon work supported in part by the National Science Foundation (NSF) under grant nos. CCF-1718549, IIS-2007887, IIS-2007951, and IIS-2143895, the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2017-17061500006, and the Defense Advanced Research Projects Agency (DARPA) and Space and Naval Warfare Systems Center Pacific (SSC Pacific) under contract no. N66001-19-C-4014.

Authors' addresses: Y. Liu, UC Santa Cruz, USA; email: yangliu@ucsc.edu; J. Wang and Y. Chen, Harvard University, USA; emails: juntaowang@g.harvard.edu, yiling@seas.harvard.edu.

2167-8375/2023/02-ART12 $15.00
https://doi.org/10.1145/3565559

how to incentivize people to provide accurate assessments has been a topic of great scientific interest.

For settings in which the principal will have access to the ground truth (e.g., after a week, knowing whether the S&P 500 index actually went up), strictly proper scoring rules (SPSRs) [4, 13, 17, 37, 46] have been developed to elicit probabilistic assessments and evaluate them against the ground truth. SPSR have two desirable properties. First, they incentivize truthful information reporting: the SPSR score of an agent's reported prediction is strictly maximized in the agent's expectation if the individual truthfully reveals one's prediction. Second, the SPSR score of a prediction measures the quality of the prediction in the sense that the closer the prediction is to the underlying, unknown true distribution of the random event, the higher the expected score.

However, in many applications, the ground truth is not available in time or at all. For example, geopolitical events usually take months to resolve [43], and whether a study will be successfully replicated is not known if a replication test of it is not attempted. In this article, we extend the literature of SPSRs to the information elicitation *without* verification (IEWV) settings, in which the principal has no access to the ground truth and still wants to elicit private probabilistic beliefs. We ask the following research question:

Can we extend SPSRs to scoring mechanisms that can achieve truthful elicitation of probabilistic information and quantify the quality of the elicited information for IEWV?

Witkowski et al. [47] explored this question in a single-task setting (i.e., having a single random variable of interest to predict). When an unbiased proxy to the true probability distribution of the ground truth is available, they generalized SPSRs to *proper proxy scoring rules*, which score a prediction against the unbiased proxy while maintaining the two properties of SPSRs. However, when the principal has access to agents' reports only, it remains an open problem as to how such an unbiased proxy can be constructed without affecting the incentive properties.

In this article, we study the research question in a multi-task setting, in which a principal wants to predict multiple random variables that are similar *a priori*. We provide a positive answer to the question. In our solution, the principal needs to know only the order of the prior probability of each possible outcome (e.g., for binary random variables, the more likely outcome) and does not need to have an unbiased proxy for each task. Specifically, we develop a family of scoring mechanisms that utilize the similarity of tasks and the conditional independence of agents' beliefs to construct a biased proxy of the ground truth. Then, we score a prediction against this proxy by removing the bias with regard to the underlying SPSRs that one wants to recover. Our proxy is explicitly constructed only from agents' reported predictions. As a result, we achieve the dominant uniform strategy truthfulness [14] in eliciting probabilistic predictions, in which truthful reporting is the strict best strategy when each agent adopts the same strategy across all tasks. Furthermore, the scores of our mechanisms recover the scores of SPSRs in expectation. To the best of our knowledge, our work provides the first meta solution that enables applications of any SPSR to the IEWV setting without relying on access to unbiased proxies of the ground truth. We name our solution *Surrogate Scoring Rules (SSRs)*.

As a building block, we first introduce SSRs for a stylized setting in which the principal has access to a noisy estimate of the ground truth, as well as the estimate's error rates, to evaluate the elicited information. We show that SSRs preserve the same information quantification and truthful elicitation properties as SPSRs despite the lack of access to the ground truth. These SSRs are inspired by the use of surrogate loss functions in machine learning [2, 5, 29, 40, 41]. They remove the bias from the noisy estimate of the ground truth such that, in expectation, a report is as if evaluated against the ground truth.

Building on this bias correction step, when the principal has access to only agents' reports and the order of the prior probabilities of each outcome, we develop the *SSR mechanisms* for the multi-task setting to achieve information quantification and the dominant uniform strategy truthfulness when the principal has sufficiently many tasks and agents. Our mechanisms rely on an estimation procedure to accurately estimate the average bias in the peer agents' reports. With the estimation, a random peer agent's report can serve as a noisy estimate of the ground truth. SSRs can then be applied to achieve the two desired properties. We evaluate the empirical performance of the SSR mechanisms using 14 real-world human forecast datasets. The results show that SSRs effectively recover SPSR scores but using only agents' reports.

We summarize our contributions as follows:

- We extend SPSRs to a family of scoring mechanisms, the SSR mechanisms, that operate in the IEWV setting. The SSR mechanisms require access only to peer reports and the order of the prior probabilities of the ground truth being each outcome, and they can truthfully elicit probabilistic beliefs. An SSR mechanism can build upon any SPSR and quantifies in expectation the value of the elicited information just as the corresponding SPSR does as if it had access to the ground truth. Therefore, our work complements the proper scoring rule literature and expands the application of SPSRs in challenging elicitation settings where the ground truth is unavailable.
- For the IEWV setting, most existing mechanisms focus on incentivizing truthful reporting of categorical signals via rewarding the correlation between two agents' reports. Our SSR mechanisms complement this literature from two perspectives. First, SSR mechanisms induce dominant uniform strategy truthfulness in eliciting probabilistic predictions instead of categorical signals. Second, instead of scoring a prediction by assessing the correlation between two agents' reports, SSR mechanisms score predictions according to their prediction accuracy against the unknown ground truth. This property encourages agents to search for more accurate forecasts.
- We evaluate the empirical performance of SSR mechanisms on 14 real-world human prediction datasets. The results show that SSR mechanisms can better reflect the true accuracy of agents in terms of SPSR scores than other existing mechanisms designed for IEWV.

Organization. The rest of the article is organized as follows. Section 2 provides a survey of related work. Section 3 introduces SPSRs and their two main desirable properties in the IEWV setting. In Section 4, we introduce our model of IEWV and our main assumptions for eliciting predictions. In Section 5, we study the information elicitation problem in the stylized setting, in which the principal has access to a noisy estimate of the ground truth with a known bias. We introduce surrogate scoring rules as a powerful solution in this section. In Section 6, we propose the dominant uniform strategy truthful mechanisms, SSR mechanisms, to address the IEWV in the multi-task binary-outcome task setting. We generalize our mechanisms and results to the multi-outcome task setting in Section 7. We present our experimental study of our mechanisms in Section 8. We discuss several restrictions of our mechanisms in Section 9. Omitted proofs can be found in the Appendix.

2 RELATED WORK

The most relevant literature to our article is on *strictly proper scoring rules* (SPSRs) and *peer prediction.* SPSRs are designed to elicit subjective beliefs about random variables when the principal can evaluate agents' predictions after the random variables are realized. Brier [4] proposed the widely used Brier score to quantify the quality of forecasts. Subsequent work studied other SPSRs and developed several characterizations of SPSRs [13, 17, 37, 46].

Peer prediction refers to a collection of mechanisms developed for incentivizing truthful reporting in IEWV. Our SSR mechanisms are additions to this collection. The core idea of peer prediction is to leverage peer reports as references to score an agent's report. The pioneer work [28] considered a single-task elicitation setting in which each agent observes a private signal associated with a single task of interest and a principal who knows the joint distribution of these signals wants to elicit the exact realizations of the signals. It proposed the first mechanism in which truthful reporting is a Bayesian Nash Equilibrium (BNE). Following this work, Jurca and Faltings [18, 19] proposed mechanisms in which truthful reporting is a BNE with a strictly higher payment than any other pure-strategy equilibrium. Kong et al. [21] proposed a mechanism, in which truthful reporting is the BNE with the highest payoff for agents among all equilibria on a binary-outcome task. Frongillo and Witkowski [11] characterized all mechanisms that admit a truthful reporting equilibrium in this setting. Another research thread for single-task elicitation asks agents to answer more questions in addition to providing their signal. The Bayesian Truth Serum [31] also asks the agents to report their beliefs about other agents' reports and then uses this additional information to score the answer of each agent. The advantage of this approach is that the principal needs not to know the joint distribution of agents' signals and that the additional information can be used to identify the correct answer to the question [32]. However, this approach introduces extra work for the agents. For interested readers, this line of research has been further developed by other studies [33, 36, 39, 49].

To relax the requirement regarding the principal's knowledge of the signal distribution, many recent peer prediction studies have focused on a multi-task setting, where there is a set of i.i.d. tasks, allowing the principal to leverage the statistical patterns in agents' reports to incentivize truthful reporting. Our work falls into this category. The multi-task setting was simultaneously developed by Dasgupta and Ghosh [6] and Witkowski and Parkes [50]. The latter were the first to explicitly estimate relevant aspects of agents' belief models from agents' reports (which our article also uses), while the former achieves provably stronger equilibrium properties. In the mechanism of Dasgupta and Ghosh [6], the truthful reporting equilibrium has the highest expected payoff for agents among all equilibria when eliciting binary signals. Radanovic et al. [35] and Shnayder et al. [42] extended the mechanism of Dasgupta and Ghosh [6] to elicit categorical signals while maintaining the same incentive property. More recent studies have achieved the dominant uniform strategy truthfulness in the multi-task setting. Parallel to our work, Kong and Schoenebeck [23] developed a framework to design mechanisms to elicit general signals as long as certain notions of mutual information can be estimated from agents' reports. Their mechanisms, which includes the mechanism of Shnayder et al. [42] as a special case, are dominant uniform strategy truthful when there is an infinite number of tasks. Kong [20] further achieved this truthfulness property with a finite number of tasks for eliciting categorical signals. Kong et al. [24] and Schoenebeck and Yu [38] proposed dominant uniform strategy truthful mechanisms to elicit continuous signals with normal distributions and with general full-support marginal distributions, respectively. When there is a noisy estimate of the ground truth with a known confusion matrix, Goel and Faltings [14] proposed a mechanism that also achieves the dominant uniform strategy truthfulness; the reward of an agent in the mechanism is an affine transformation of the the agent's correctness rate over all classes. In comparison, our dominant uniform strategy truthfulness mechanisms focus on eliciting posterior beliefs of the ground truth and the rewards in our mechanisms recover in expectation the accuracy of agents in terms of the SPSR. Instead of assuming availability of an estimate of the confusion matrix, we construct an estimate from the agents' reports, assuming that the principal knows the order of the prior probabilities of each possible outcome of the ground truth.

There are a few studies also focusing on eliciting probabilistic predictions, as we are doing in our article. Among these studies, Witkowski and Parkes [48] and Radanovic and Faltings [34] consider single-task elicitation and ask agents to report additional information as required by the Bayesian Truth Serum [31]. The two mechanisms proposed make truthful reporting an ex-post equilibrium and a BNE, respectively. Kong and Schoenebeck [22] provided a mechanism to elicit probabilistic predictions for the multi-task setting. Although truthful reporting is an equilibrium strategy under their mechanism, the mechanism is not dominant uniform strategy truthful. When the principal has access to an unbiased proxy of the ground truth, the proxy scoring rules developed by Witkowski et al. [47] can be used to elicit probabilistic predictions for the single-task setting as to what SPSRs offer with access to the ground truth. In this case, proxy scoring rules score a prediction against the unbiased proxy using an SPSR, and the expected score is equal to the expected score given by the SPSR using the ground truth up to a positive affine transformation [8]. In comparison, our mechanisms also offer a meta approach to recover the score for any SPSR. Our mechanisms do not require access to an unbiased proxy; rather, they require a set of i.i.d. tasks.

Finally, our work borrows ideas from the machine learning literature on learning with noisy data [9, 29, 40, 44]. At a high level, our goal in this article aligns with the goal in learning from noisy labels – both aim to evaluate a prediction when the ground truth is missing but, instead, a noisy signal of the ground truth is available. Our work addresses the additional challenge that the error rate of the noisy signal remains unknown *a priori*.

3 PRELIMINARIES

Before we introduce our model of IEWV, we first briefly introduce SPSRs, which are designed for the well-studied IEWV settings. We highlight two useful properties of SPSRs: (1) SPSRs quantify the value of information and (2) SPSRs are incentive compatible for elicitation. Our goal of this article is to develop scoring rules that match these properties for the more challenging without-verification settings. Our solutions build upon the understanding of SPSRs.

SPSRs are designed for eliciting subjective distributions of random variables when the principal can reward agents after the realization of the random variables. SPSRs apply to eliciting predictions for any random variables. However, we introduce them for binary random variables in this section because the rest of our article focuses on the binary case. Let $Y \in \{0, 1\}$ represent a binary event. An agent has a subjective belief $p \in [0, 1]$ for the likelihood of $Y = 1$. When the agent reports a probabilistic prediction $q \in [0, 1]$ of Y being 1, the principal rewards the agent using a scoring function $S(q, y)$ that depends on both the agent's report q and the realized outcome of Y. Strict properness of $S(\cdot, \cdot)$ is defined as follows.

Definition 3.1. A function $S : [0, 1] \times \{0, 1\} \to \mathbb{R}$ that maps a reported belief q and the ground truth Y into a score is a *strictly proper scoring rule* if it satisfies $\underline{\mathbb{E}[S(p, Y)] > \mathbb{E}[S(q, Y)]}$, for all $p, q \in [0, 1]$ and $p \neq q$. Both expectations are taken with respect to $Y \sim \text{Bernoulli}(p)$.

There is a rich family of strictly proper scoring rules, including the Brier score ($S(q, Y) = 1 - (q - Y)^2$), the log scoring rule ($S(q, Y) = \log(q)$ if $Y = 1$ and $S(q, Y) = \log(1 - q)$ if $Y = 0$) and the spherical scoring rules [13].

Incentive compatibility of SPSRs. The definition of SPSRs immediately gives incentive compatibility. If an agent's belief is p, reporting p truthfully uniquely maximizes the agent's expected score.

SPSRs quantify the value of information. Another useful property of SPSRs is that they quantify the value/accuracy of reported predictions. To present a rigorous argument, we use an indicator vector $\mathbf{y}$ of length 2 to represent the realization of Y, with 1 at the Y-th position and 0 otherwise. That is, $\mathbf{y} = (0, 1)$ if $Y = 1$ and $\mathbf{y} = (1, 0)$ if $Y = 0$. We use a probability vector $\mathbf{q} = (1 - q, q)$ to represent probability q. By the representation theorem [13, 26, 37], any SPSR can be characterized using a corresponding strictly convex function G as follows: $S(\mathbf{q}, \mathbf{y}) = G(\mathbf{y}) - D_G(\mathbf{y}, \mathbf{q})$, where D_G is the Bregman divergence function of G. Now, consider the unknown true distribution of Y, denoted by $\mathbf{p}^* = (1 - p^*, p^*)$. The expected score for an agent predicting $\mathbf{q}$ is

$$\mathbb{E}[S(\mathbf{q}, \mathbf{y})] = \mathbb{E}[G(\mathbf{y})] - \mathbb{E}[D_G(\mathbf{y}, \mathbf{q})],$$

where all three expectations are taken over $Y \sim \text{Bernoulli}(p^*)$. This means that the maximum score an agent can receive in expectation is $\mathbb{E}_{Y\sim\text{Bernoulli}(p^*)}[G(\mathbf{y})]$, which happens when the agent's report $\mathbf{q} = \mathbf{p}^*$. Moreover, a prediction $\mathbf{q}$ with a smaller divergence $\mathbb{E}_{y\sim\mathbf{p}^*}[D_G(\mathbf{y}, \mathbf{q})]$ receives a higher score in expectation. Intuitively, $\mathbb{E}_{Y\sim\text{Bernoulli}(p^*)}[D_G(\mathbf{y}, \mathbf{q})]$ characterizes how "far away" $\mathbf{q}$ is from the true distribution of Y under divergence function D_G. This implies that a strictly proper scoring rule S qualifies the accuracy of a prediction $\mathbf{q}$ based on the corresponding divergence function. When S is taken as the Brier scoring rule, the corresponding Bregman divergence is the quadratic function, and $\mathbb{E}_{Y\sim\text{Bernoulli}(p^*)}[D_G(\mathbf{y}, \mathbf{q})] = ||\mathbf{p}^* - \mathbf{q}||^2$, implying that a prediction $\mathbf{q}$ closer to $\mathbf{p}^*$ according to ℓ_2 norm receives a higher score in expectation. When S is taken as the log scoring rule, the corresponding Bregman divergence is the KL-divergence, D_{KL}, which is also called the relative entropy, and $\mathbb{E}_{Y\sim\text{Bernoulli}(p^*)}[D_G(\mathbf{y}, \mathbf{q})] = D_{KL}(\mathbf{p}^*||\mathbf{q}) + H(\mathbf{p}^*)$, where H is the entropy function. A prediction with a smaller KL-divergence from $\mathbf{p}^*$ receives a higher score in expectation. This property of SPSRs allows the principal to take an expert's average score over a set of prediction tasks as a proxy of the expert's average accuracy and rank experts accordingly.

4 MODEL AND MECHANISM DESIGN PROBLEM

We consider a multi-task setting for the IEWV problem Under this setting, we aim to develop scoring mechanisms that are incentive compatible and are able to quantify the value of elicited information, recovering the two desirable properties that SPSRs achieve in the presence of the ground truth. In this section, we formally introduce the information structure of our setting and the exact mechanism design problem that we consider.

4.1 Model of Information Structure

A principal has a set of tasks $[m] = \{0, \dots, m - 1\}$. Each task asks for a prediction for an independent random variable of interest, denoted by Y_k, $k \in [m]$. For now, we assume that these random variables to predict are binary variables, that is, $Y_k \in \{0, 1\}, \forall k \in [m]$. We will generalize our results to (non-binary) categorical random variables in Section 7. There is a set of informed agents $[n] = \{0, \dots, n - 1\}$. Each agent $i \in [n]$ privately observes a random signal $O_{i,k}$ generated by Y_k for each task $k \in [m]$ and, thus, holds a posterior belief about Y_k, represented by $P_{i,k} := \Pr[Y_k = 1|O_{i,k}]$. The posterior $P_{i,k}$ is a random variable as the signal $O_{i,k}$ is a random variable. We make the following main assumptions on the information structure among the signals and ground truth.

Assumption 1. *Tasks are independent and similar* a priori, *that is, the joint distribution of* $(O_{1,k}, \dots, O_{n,k}, Y_k)$ *is i.i.d. for all tasks* $k \in [m]$.

This assumption is natural when the set of tasks are of similar nature, for example, tasks to predict the replicability of multiple studies published in the same journal and the same year. In this example, readers may *a priori* hold the same journal-wide belief about the features and replicability of each study. After reading the journal, each agent receives a private signal for each individual study, which allows the agent to provide a more informed prediction for that study. This assumption is common for multi-task IEWV.[1]

[1]Kong [20], Kong and Schoenebeck [22] consider information elicitation for objective questions (i.e., questions for which an objective ground truth exists). They make the same assumption as Assumption 1. Other studies (e.g., [6, 20, 23, 35, 42])

Based on Assumption 1, each Y_k has the same prior, denoted by $p := \Pr[Y_k = 1]$. Also, for a fixed agent i, the distribution of signal $O_{i,k}$ conditioned on Y_k on each task $k \in [m]$ is the same. We use $\mathcal{D}_i^+$ and $\mathcal{D}_i^-$ to denote this conditional distribution for agent i for conditions $Y_k = 1$ and $Y_k = 0$, respectively. We assume that $\mathcal{D}_i^+ \neq \mathcal{D}_i^-$; otherwise, observation $O_{i,k}$ is independent and uninformative to Y_k. Each agent forms a posterior belief $P_{i,k}$ using the prior p and the conditional distributions $\mathcal{D}_i^+$ and $\mathcal{D}_i^-$. We require no knowledge of $\mathcal{D}_i^+$ and $\mathcal{D}_i^-$ for the principal and the agents other than agent i. We assume that agents' signals are independent conditioned on the ground truth.

Assumption 2. *For each task, agents' signals are mutually independent conditioned on the ground truth, that is,* $\forall k \in [m], \Pr\left[O_{1,k}, \ldots, O_{n,k} | Y_k\right] = \prod_{i \in [n]} \Pr[O_{i,k} | Y_k]$.

This assumption excludes the scenarios in which agents have some form of "side information" to coordinate their reports. With "side information," it is impossible to have any mechanism that can truthfully elicit agents' predictions without access to ground truth. This issue has been noted by Kong and Schoenebeck [22] and Kong [20] for tasks with ground truth and the same assumption has been adopted. Finally, we make a technical assumption about the prior p and the principal's knowledge.

Assumption 3. *It holds that* $p \neq 0.5$ *and the principal knows whether* $p > 0.5$ *or not.*

We do not assume that the principal knows the exact prior p of tasks but assume that the principal knows whether $p > 0.5$ or $p < 0.5$. This one binary-bit of information helps the principal distinguish between the set of truthful predictions and the set of inverted predictions (i.e., everyone reporting $1 - p_{i,k}$ instead of $p_{i,k}$), which otherwise is impossible to distinguish. In practice, this information is usually easy to obtain. In the example of predicting the replicability of studies, this assumption requires only that the principal knows whether the majority of the studies can be replicated or not. The assumption $p \neq 0.5$ is a technical condition we need in order to distinguish the truthful reporting scenario from the inverted reporting scenario.

We also assume that the posterior $P_{i,k}$ for any agent i on any task k is different under different realizations of private signal $O_{i,k}$. This assumption is without loss of generality because different realizations of $O_{i,k}$ that lead to the same posterior $P_{i,k}$ for agent i on task k also lead to the same posterior about any other agent's signal $O_{j,k}$ for agent i due to Assumption 2. Therefore, we can merge multiple realizations of $O_{i,k}$ that lead to the same posterior $P_{i,k}$ into one realization without influencing agent i's belief about other agents' signals and the ground truth. Consequently, it is without loss of generality to assume that there exists a one-to-one correspondence between the realization of an agent's signal $O_{i,k}$ and the agent's posterior $P_{i,k}$. According to this one-to-one correspondence and Assumptions 1 and 2, the following two conditions hold for $P_{i,k}$ for $i \in [n], k \in [m]$.

Proposition 4.1. *Under Assumptions 1 and 2, the following two conditions hold for agents' beliefs* $P_{i,k}, i \in [n], k \in [m]$.

(1) $P_{1,k}, \ldots, P_{n,k}$ *and* Y_k *are independent of their own counterparts across tasks* $k \in [m]$ *but have the same joint distribution, that is,* $(P_{1,k}, \ldots, P_{n,k}, Y_k)$ *are i.i.d. across tasks* $k \in [m]$.
(2) For each task $k \in [m]$, $P_{1,k}, \ldots, P_{n,k}$ *are independent conditioned on* Y_k, *that is,* $\Pr[P_{1,k}, \ldots, P_{n,k} | Y_k] = \prod_{i \in [n]} \Pr[P_{i,k} | Y_k], \forall k \in [m]$.

The first condition in Proposition 4.1 implies that an agent has the same expertise level across different tasks, as the joint distribution of the agent's posterior belief and the ground truth is

consider information elicitation for subjective questions (i.e., questions with no objective ground truth, e.g., how do you rate the movie?). These studies also assume that the joint distributions of agents' signals are the same across all tasks.

the same across tasks. The second condition implies that given the ground truth, each agent's probabilistic prediction is independent. The two conditions in Proposition 4.1 in fact characterize a broader space of information structure than the space captured by Assumptions 1 and 2. The former space includes the information structure in which each task has a different prior but the distribution of the posterior beliefs of each agent are still the same across tasks. Our theoretical results in this article hold for the model with this broader information structure space characterized by the two conditions in Proposition 4.1 and with Assumption 3, where p refers to the mean prior overall tasks.

4.2 Mechanism Design Problem

We consider the multi-task peer prediction mechanisms in which the principal assigns each task k to a subset $[n_k] \subseteq [n]$ of agents, collects a single probabilistic prediction $q_{i,k} \in [0,1]$ from each agent i assigned with task k, and pays each agent based on all predictions collected from all agents. We use $[m_i] \subseteq m$ to denote the set of tasks assigned to agent i. We use $q_{i,k} = \emptyset$ to denote that agent i has not been assigned to task k. Such a multi-task peer prediction mechanism can be formally expressed as a function $R : \{\emptyset \cup [0,1]\}^{n\times m} \to \mathbb{R}^n$, which maps a prediction profile on all tasks and all agents to a vector of total payments of all agents. In this article, we restrict our attention to anonymous mechanisms that give each prediction from an agent an independent payment such as an SPSR. Thus, a mechanism that we consider can be fully expressed by a score function $R : \{\emptyset \cup [0,1]\} \times \{\emptyset \cup [0,1]\}^{n-1\times m} \to \mathbb{R}$, which maps a single prediction $q_{i,k}$ of agent i on task k and a profile of predictions from all other agents into a single reward score for that prediction, and agent i's total reward is the sum of the scores that the agent obtains across the tasks the agent is assigned.

Agents have no obligation to report their true beliefs. Instead, given a mechanism, an agent can report strategically to maximize one's expected payment. As there is a one-to-one correspondence between an agent's signal and the individual's posterior on a single task in our model, we can define an agent's reporting strategy on a single task without loss of generality as a function that maps the agent's posterior to a distribution where that individual's reported prediction is drawn from.

Definition 4.2. Let $\Delta_{[0,1]}$ be the space of all probability distributions over $[0,1]$. The strategy of an agent i on task k is a mapping $\sigma_i : [0,1] \to \Delta_{[0,1]}$, which maps the agent's posterior belief $P_{i,k}$ into a distribution $\sigma_i(P_{i,k})$ over [0,1], from which the agent draws the reported prediction $Q_{i,k}$.

We use the uppercase $Q_{i,k}$ to denote the reported prediction when we want to emphasize that the reported prediction is a random variable determined by an agent's posterior belief and that individual's reporting strategy jointly; otherwise, $q_{i,k}$ is used. We further assume that each agent adopts the same mixed strategy across all asigned tasks.

Assumption 4 (Uniform Strategy). *For any agent $i \in [n]$, the agent adopts the same strategy $\sigma_i(\cdot)$ over all assigned tasks $k \in [m_i]$.*

This assumption is reasonable as we assume that tasks are *a priori* similar to each other. We use $\sigma_i(\cdot)$ to denote the reporting strategy adopted by agent i on all tasks that the agent answers and use σ_{-i} to denote the strategy profile used by all agents except agent i. We use $\mathbb{E}[R(q_{i,k};\sigma_{-i})]$ to denote the expected score that agent i receives for reporting $q_{i,k}$ when other agents use strategy profile σ_{-i}, where the expectation is taken over the randomness in ground truth, other agents' signals and strategies, and in the mechanism itself. We use $\mathbb{E}[R(\sigma_i;\sigma_{-i})]$ to denote the expected reward of agent i when the agent's report is also a random variable generated by that individual's belief $P_{i,k}$ and reporting strategy σ_i.

In this article, our goal is to design a mechanism $R(\cdot)$ in the IEWV setting with similar properties that SPSRs have for the IEWV setting: *quantification of the value of information* and *incentive compatibility*.

Quantifying value of information. The score of each prediction should reflect the true accuracy of the prediction, similar to what SPSRs achieve. That is, for all i, k, and $q_{i,k}$ and for any true distribution of the ground truth Y_k, $\underline{\mathbb{E}[R(q_{i,k};\sigma_{-i})] = f\left(E_{Y_k}[S(q_{i,k}, Y_k)]\right)}$ holds for an SPSR $S(\cdot)$ and a strictly increasing function f, where the two expectations are taken over the true distributions of the random variables in the two expressions at each side of the equality. This design goal pursues that the score that an agent receives for a prediction in IEWV recovers what the agent would receive with an SPSR (with access to the ground truth) in expectation.

Incentive Compatibility. A mechanism satisfies incentive compatibility to some extent if truthful reporting is a strategy that maximizes an agent's expected utility under certain conditions. In this article, we pursue the dominant uniform strategy truthfulness [14], where truthful reporting is a dominant strategy if we restrict the strategy space with the uniform strategy assumption (Assumption 4).

Formally, in IEVW, a dominant uniform strategy truthful mechanism is a mechanism in which truthful reporting on each task maximizes an agent's expected reward no matter what uniform strategies the other agents play and strictly maximizes the agent's expected reward if other agents' reports are also informative.[2] Let σ_i^* be the truthful reporting strategy for agent i, that is, σ_i^* is the function that maps a belief p_i to a distribution in which all probability mass is put on p_i. Let $\underline{\bar{Q}_{-i,k} := \frac{1}{n-1}\sum_{j\neq i} Q_{j,k}}$ be the mean of all agents' reported predictions on task k except agent i's. Note that $\bar{Q}_{-i,k}$ is a random variable because of the randomness in reporting strategy σ_j and the randomness in signal $O_{j,k}$ for all $j \neq i$. We say that $\bar{Q}_{-i,k}$ is informative about the ground truth if $\mathbb{E}[\bar{Q}_{-i,k}|y_k = 1] \neq \mathbb{E}[\bar{Q}_{-i,k}|y_k = 0]$. We formally define the dominant uniform strategy truthful mechanisms as follows.

Definition 4.3 (Dominant Uniform Strategy Truthfulness). A mechanism $R(\cdot)$ is *dominant uniform strategy truthful* if $\forall i \in [n]$, $\forall k \in [m_i]$, $\forall\{\mathcal{D}_j^+, \mathcal{D}_j^-\}_{j\in[n]}$ and for any realization $o_{i,k}$ of signal $O_{i,k}$: $\mathbb{E}[R(\sigma_i^*;\sigma_{-i})|O_{i,k} = o_{i,k}] \geq \mathbb{E}[R(\sigma_i;\sigma_{-i})|O_{i,k} = o_{i,k}]$ for any uniform strategy $\sigma_i \neq \sigma_i^*$ and any uniform strategy profile of other agents σ_{-i}, and the inequality holds strictly for any uniform strategy profile σ_{-i} under which $\bar{Q}_{-i,k}$ is informative about Y_k.

In Definition 4.3, we characterize the condition that peers' reports are informative as follows: that the expectation of the mean of peers' reports on a task differs when conditioned on different realizations of the ground truth.

5 ELICITATION WITH A NOISY ESTIMATE OF GROUND TRUTH

Before we develop mechanisms with the two desirable properties we pursue, in this section we first obtain these two properties under a very stylized setting: *elicitation with a noisy estimate of*

[2]In a standard dominant truthful mechanism, truthful reporting strictly maximizes the agent's expected reward no matter what strategies other agents play. In IEWV, however, if all peer agents report predictions independently with regard to the ground truth on each task, then there will be no information available for the mechanism to incentivize truthful reporting. Therefore, it is inevitable to allow a dominant truthful mechanism in IEWV to pay truthful reporting strictly more only when the peer reports are informative about the ground truth. For example, in studies [20, 23], the dominant uniform strategy truthful mechanism is defined to be a mechanism that pays truthful reporting strictly more only when for each agent, there is at least one peer agent reporting truthfully. We will see later that in our definition, we do not require that there is at least one peer agent reporting truthfully. We allow all peer agents to play non-truthfully, but require the mean of peer agents' reports to be informative with respective to the ground truth.

ground truth. In this stylized setting, we introduce surrogate scoring rules as an effective solution. These scoring rules will be the building blocks of our mechanisms designed for the general setting.

This stylized setting has only one event Y and one agent i, who observes a signal O_i generated from distribution $\mathcal{D}_i(Y)$ and forms a posterior $P_i = \Pr[Y = 1|O_i]$. The principal in this setting has access to a noisy estimate $Z \in \{0, 1\}$ of the ground truth Y, although the individual has no access to the exact realization of Y. The noisy estimate Z is characterized by two *error rates*, e_z^+ and e_z^-, defined as $e_z^+ := \Pr[Z = 0|Y = 1]$, $e_z^- := \Pr[Z = 1|Y = 0]$, which are the probabilities that Z mismatches Y under the two realizations of Y. The principal knows the realization Z and the exact error rates e_z^+, e_z^-. The principal cannot expect to do much if Z is independent of Y. Therefore, we assume that Z and Y are stochastically relevant, an assumption commonly adopted on the relation between a signal and the ground truth in the information elicitation literature [28].

Definition 5.1. A random variable Z is *stochastically relevant* to a random variable Y if the distribution of Y conditioned on Z differs for different realizations of Z.

The following lemma shows that the stochastic relevance condition directly translates to a constraint on the error rates, that is, $e_z^+ + e_z^- \neq 1$. This lemma can be proved immediately by writing out the distribution of Y conditioned on Z in terms of the two error rates e_z^+, e_z^- and the prior of Z.

Lemma 5.2. *The noisy estimate Z is stochastically relevant to the ground truth Y if and only if $e_z^+ + e_z^- \neq 1$.*

The goal of the principal in this setting is to design a scoring rule to elicit the posterior P_i truthfully based on this noisy estimate Z and the error rates e_z^+, e_z^-. We define the design space of the scoring rules with the noisy estimate as follows.

Definition 5.3. Given a noisy estimate Z of ground truth Y with error rates $(e_z^+, e_z^-) \in [0, 1]^2$, *a scoring rule against the noisy estimate of the ground truth* is a function $R : [0, 1] \times \{0, 1\} \to \mathbb{R}$ that maps a prediction $q_i \in [0, 1]$ and a realized noisy estimate $z \in \{0, 1\}$ to a score. The function R can depend on the two error rates (e_z^+, e_z^-).

Adopting the terminology from the scoring rule literature, we refer to strict properness of a scoring rule against a noisy estimate of ground truth as the property that the rule assigns a strictly better expected score to a truthful prediction of the ground truth than to a non-truthful prediction.

Definition 5.4. A scoring rule $R(q_i, Z)$ against a noisy estimate Z of ground truth is *strictly proper* for eliciting an agent's posterior belief generated by signal O_i if it holds for all realizations o_i of O_i and the posterior $p_i = \Pr[Y = 1|O_i = o_i]$ that

$$\mathbb{E}_{Z|O_i=o_i}[R(p_i, Z)] > \mathbb{E}_{Z|O_i=o_i}[R(q_i, Z)], \forall q_i \in [0, 1](q_i \neq p_i).$$

5.1 Surrogate Scoring Rules (SSRs)

In this section, we present our solution, the *surrogate scoring rules* (SSRs), for this stylized setting. SSRs are a family of scoring rules that evaluate a prediction against a noisy estimate of ground truth. For any distribution of the ground truth and any stochastically relevant noisy estimate of the ground truth, the expected score that SSRs give to the prediction, with expectation taken over the randomness of the noisy estimate, is equal to (up to a monotonic increasing transformation) the expected score that an SPSR gives to the same prediction, with expectation taken over the randomness of the ground truth. We will see that SSRs are strictly proper under mild conditions.

Definition 5.5 (Surrogate Scoring Rules). $R : [0, 1] \times \{0, 1\} \to \mathbb{R}$ is a surrogate scoring rule if for some strictly proper scoring rule $S : [0, 1] \times \{0, 1\} \to \mathbb{R}$ and a strictly increasing function $f : \mathbb{R} \to \mathbb{R}$, it holds that $\forall p_i, q_i, e_z^+, e_z^- \in [0,1]$ and $e_z^+ + e_z^- \neq 1$, $\underline{\mathbb{E}_Z[R(q_i, Z)] = f(\mathbb{E}_Y[S(q_i, Y)])}$, where Y is the ground truth drawn from Bernoulli(p_i) and Z is a noisy estimate of Y with error rates e_z^+, e_z^-.

Definition 5.5 defines the SSR $R(\cdot)$ as scoring rules that help us remove the bias in Z and return us the same score given by an SPSR in expectation. The idea of SSRs is borrowed from the machine learning literature on learning with noisy data [5, 27, 29, 40, 44]. SSRs can be viewed as a particular class of the proxy scoring rules proposed by Witkowski et al. [47]. Witkowski et al. [47] achieve properness of proxy scoring rules by plugging in an *unbiased* proxy of the ground truth to an SPSR. With SSRs, we directly work with biased proxy and design scoring functions to de-bias the noise in the proxy. We have the following strict properness result for SSRs straightforwardly:

Theorem 5.6. *Given the prior p of the ground truth Y and a private signal O_i, SSR $R(q_i, Z)$ against a noisy estimate Z is strictly proper for eliciting the posterior $P_i = \Pr[Y = 1|O_i]$ if Z and O_i are independent conditioned on Y and Z is stochastically relevant to Y.*

We provide an implementation of SSRs, which we call SSR$_\alpha$:

$$R(q_i, Z = 1) = \frac{(1 - e_z^-) \cdot S(q_i, 1) - e_z^+ \cdot S(q_i, 0)}{1 - e_z^+ - e_z^-}, \tag{1}$$

$$R(q_i, Z = 0) = \frac{(1 - e_z^+) \cdot S(q_i, 0) - e_z^- \cdot S(q_i, 1)}{1 - e_z^+ - e_z^-}, \tag{2}$$

where S can be any strictly proper scoring rule. This SSR implementation is inspired by Natarajan et al. [29]. As can been seen from Equations (1) and (2), the knowledge of the error rates e_z^+, e_z^- is crucial for defining SSR$_\alpha$. Moreover, SSR$_\alpha$ has the property that the expected score $\mathbb{E}_{Z|Y}[R(q_i, Z)]$ conditioned on the realization of the ground truth Y is exactly the same as the score $S(q_i, Y)$ given by the SPSR. More formally, we have the following lemma.

Lemma 5.7 (Lemma 1, [29]). *For SSR$_\alpha$, ground truth Y and noisy estimate Z, $\forall q_i, e_z^+, e_z^- \in [0,1]$ and $e_z^+ + e_z^- \neq 1, \forall y \in \{0,1\} : \mathbb{E}_{Z|Y=y}[R(q_i, Z)] = S(q_i, Y = y)$.*

Proof. Lemma 1 in [29] proves the statement for $e_z^+ + e_z^- < 1$. For completeness, we provide the proof for $e_z^+ + e_z^- \neq 1$ here. Let $q_i \in [0,1]$ be an arbitrary prediction. When $y = 1$, we have that

$$\begin{aligned}
\mathbb{E}_{Z|Y=1}[R(q_i, Z)] &= (1 - e_z^+)R(q_i, 1) + e_z^+ R(q_i, 0) \\
&= (1 - e_z^+)\frac{(1 - e_z^-)S(q_i, 1) - e_z^+ S(q_i, 0)}{1 - e_z^+ - e_z^-} + e_z^+ \frac{(1 - e_z^+)S(q_i, 0) - e_z^- S(q_i, 1)}{1 - e_z^+ - e_z^-} \\
&= \frac{\left((1 - e_z^+)(1 - e_z^-) - e_z^+ e_z^-\right) S(q_i, 1)}{1 - e_z^+ - e_z^-} \\
&= S(q_i, 1).
\end{aligned}$$

When $y = 0$, we have that

$$\begin{aligned}
\mathbb{E}_{Z|Y=0}[R(q_i, Z)] &= e_z^- R(q_i, 1) + (1 - e_z^-)R(q_i, 0) \\
&= e_z^- \frac{(1 - e_z^-)S(q_i, 1) - e_z^+ S(q_i, 0)}{1 - e_z^+ - e_z^-} + (1 - e_z^-)\frac{(1 - e_z^+)S(q_i, 0) - e_z^- S(q_i, 1)}{1 - e_z^+ - e_z^-} \\
&= S(q_i, 0).
\end{aligned}$$

□

Intuitively, the linear transformation in SSR$_\alpha$ ensures that, in expectation, the prediction q_i is scored as if it were scored against the ground truth Y under the underlying SPSR. We would like to note that other surrogate loss functions designed for learning with noisy labels can also be leveraged to design SSRs. With the conditional unbiasedness property of SSR$_\alpha$, we can formally claim that SSR$_\alpha$ is a surrogate scoring rule, as stated in Theorem 5.8 next.

Theorem 5.8. *SSR_α is a surrogate scoring rule and $\forall p_i, q_i, e_z^+, e_z^- \in [0,1] (e_z^+ + e_z^- \neq 1)$, $\mathbb{E}_Z[R(q_i, Z)] = \mathbb{E}_Y[S(q_i, Y)]$, where Y is the ground truth drawn from Bernoulli(p_i) and Z is the noisy estimate of ground truth Y with error rate e_z^+, e_z^-.*

Proof. As shown by Lemma 5.7, for SSR_α, we have that $\forall p_i, q_i, e_z^+, e_z^-$ $(e_z^+ + e_z^- \neq 1)$ and $\forall y \in \{0,1\}$, $\mathbb{E}_{Z|Y=y}[R(q_i, z)] = S(q_i, Y = y)$, we have immediately that

$$\mathbb{E}_Z[R(q_i, z)] = \mathbb{E}_Y[\mathbb{E}_{Z|Y}[R(q_i, Z)]] = \mathbb{E}_Y[S(q_i, Y)]. \qquad \square$$

With Theorem 5.8, we know that SSR_α quantifies the quality of information of a prediction just as the underlying SPSR S does. Furthermore, SSR_α has the following variance:

Theorem 5.9. *Let $p_z := \Pr[Z = 1]$. For a fixed prediction $q_i \in [0,1]$, SSR_α suffers the following variance:*

$$\mathbb{E}_Z\Big[R(q_i, Z) - \mathbb{E}_z[R(q_i, Z)]\Big]^2 = \frac{p_z \cdot (1 - p_z)}{(1 - e_z^+ - e_z^-)^2} \cdot (S(q_i, 1) - S(q_i, 0))^2 \ . \tag{3}$$

Proof.

$$\begin{aligned}
&\mathbb{E}_Z\Big[R(q_i, Z) - \mathbb{E}_Z[R(q_i, Z)]\Big]^2 \\
&= p_z\Big(R(q_i, 1) - \big(p_z R(q_i, 1) + (1 - q_z) R(q_i, 0)\big)\Big)^2 \\
&\quad + (1 - p_z)\Big(R(q_i, 0) - \big(p_z R(q_i, 1) + (1 - q_z) R(q_i, 0)\big)\Big)^2 \\
&= p_z (1 - p_z)^2 \big(R(q_i, 1) - R(q_i, 0)\big)^2 + (1 - p_z) p_z^2 \big(R(q_i, 0) - R(q_i, 1)\big)^2 \\
&= p_z (1 - p_z) \big(R(q_i, 0) - R(q_i, 1)\big)^2 \\
&= \frac{p_z (1 - p_z)}{(1 - e_z^+ - e_z^-)^2} \Big((1 - e_z^-) S(q_i, 1) - e_z^+ S(q_i, 0) - \big((1 - e_z^+) S(q_i, 0) - e_z^- S(q_i, 1)\big)\Big)^2 \\
&= \frac{p_z (1 - p_z)}{(1 - e_z^+ - e_z^-)^2} \big(S(q_i, 1) - S(q_i, 0)\big)^2. \qquad \square
\end{aligned}$$

6 ELICITATION WITHOUT VERIFICATION

The results in the previous section are built upon the fact that there is a noisy estimate of ground truth with known error rates. In this section, we apply the idea of SSRs to the IEWV setting. A reasonable way to do so is to use agents' reports as the source of the noisy estimate. Although the principal does not know the exact bias in agents' reports, we find a way to construct such a noisy proxy of ground truth and estimate its error rates. We refer to this noisy proxy as the *reference report*. Applying SSRs with this reference report, we can finally get a family of mechanisms that are dominant uniform strategy truthful and that also quantify the value of information in agents' reports as to what SPSRs do. Within this family, we can choose different underlying SPSRs for SSRs to get different mechanisms. We call this family of mechanisms *SSR mechanisms*. We present a sketch of our SSR mechanisms in Mechanism 1.

The challenge of designing such mechanisms is to construct the reference report $Z_{i,k}$ in Mechanism 1 and successfully estimate its error rates $e^+_{z_{i,k}}, e^-_{z_{i,k}}$. In the following sections, we show how to construct this reference report and estimate its error rates.

Mechanism 1: SSR mechanisms (Sketch)

1: For each task k, we uniformly randomly pick at least 3 agents, assign task k to them, and collect their predictions.
2: For each agent i and each task k that the agent answers, we construct a reference report $Z_{i,k}$ using the agent's peer agents' reports, and estimate the error rates $e^+_{z_{i,k}}$ and $e^-_{z_{i,k}}$ for $Z_{i,k}$.
3: Pay each agent i for the agent's prediction $q_{i,k}$ on task k by SSR $R(q_{i,k}, Z_{i,k})$ if $e^+_{z_{i,k}} + e^+_{z_{i,k}} \neq 1$, and pay 0, otherwise.

6.1 Reference Report and Its Property

Recall that we use $Q_{j,k}$ to denote the reported prediction of agent j on task k, which is generated by agent j's posterior belief $P_{j,k}$ and reporting strategy σ_j. Let $S_{j,k} \in \{0, 1\}$ be a binary signal independently drawn from Bernoulli($Q_{j,k}$). We refer to $S_{j,k}$ as the *prediction signal* of agent j on task k. We construct the reference report $Z_{i,k}$ for agent i as follows: *We uniformly randomly pick an agent j from agent i's peer agent set $[n]\backslash\{i\}$, collect agent j's prediction $Q_{j,k}$, and draw a prediction signal $S_{j,k} \sim$ Bernoulli($Q_{j,k}$). We use this $S_{j,k}$ as the reference report $Z_{i,k}$ for agent i on task k.*

Conditioned on all peer agents' reports $Q_{j,k}, j \in [n]\backslash\{i\}$, the distribution of $Z_{i,k}$ is Bernoulli($\bar{Q}_{-i,k}$), because we pick a prediction signal from all peer agents uniformly randomly. Recall that in our model, $Q_{i,k} \sim \sigma_i(P_{i,k}), i \in [n], k \in [m]$. Due to Proposition 4.1 and Assumption 4, $\bar{Q}_{-i,k}$ is i.i.d. across tasks $k \in [m]$ and is independent to agent i's posterior $P_{i,k}$ conditioned on the ground truth Y_k for any task k. Therefore, $Z_{i,k}, k \in [m]$ that we construct have the following two properties.

Lemma 6.1. *$\forall i \in [n], k \in [m]$, $Z_{i,k}$ is independent to agent i's posterior $P_{i,k}$ conditioned on Y_k.*

This property ensures that $Z_{i,k}$ can be used as the conditionally independent noisy estimate of the ground truth in Theorem 5.6 and, thus, SSR against $Z_{i,k}$ is strictly proper for eliciting the posterior belief $P_{i,k}$.

Lemma 6.2. *For any strategy profile agents play, the reference reports of a single agent i for any $i \in [n]$ are i.i.d. across tasks and have the same error rates with regard to their corresponding ground truth Y_k, that is, $\forall \sigma_1, \ldots, \sigma_n, \forall i \in [n], \exists e^+_i, e^-_i \in [0, 1], \forall k \in [m]$: $\Pr[Z_{i,k} = 0|Y_k = 1] = e^+_i, \Pr[Z_{i,k} = 1|Y_k = 0] = e^-_i$.*

This lemma shows that the error rates of the reference reports of an agent i are the same across all tasks. This property allows the estimation of the error rates using the multi-task prediction data. In the following sections, we introduce the estimation of the error rates and complete our mechanisms. We prove Lemmas 6.1 and 6.2 below.

Proof. Proposition 4.1 and Assumption 4 directly imply that (1) for each task, $Q_{1,k}, \ldots, Q_{n,k}$ are mutually independent conditioned on the ground truth Y_k, and (2) $(Q_{1,k}, \ldots, Q_{n,k}, y_k)$ are i.i.d across tasks $k \in [M]$. As $Z_{i,k}$ is independently drawn from Bernoulli($\bar{Q}_{-i,k}$), we immediately have that (1') for each task $k \in [m]$, $Z_{i,k}$ is independent to $O_{i,k}$ and, thus, to $P_{i,k} := \Pr[Y_k = 1|O_{i,k}]$, and (2') $(Z_{i,k}, Y_k)$ have the same joint distribution for $k \in [m]$. As a result of (2'), $Z_{i,k}, k \in [m]$ have the same error rates with regard to the corresponding Y_k. □

6.2 Asymptotic Setting

To better deliver our idea for error rate estimation, we start with an asymptotic setting with infinite amounts of tasks and agents, that is, $m, n \to \infty$. We will later provide a finite sample justification for our mechanism. Based on Lemma 6.2, the reference reports of an agent on different tasks have the same distribution and error rates. Therefore, we focus on estimating the error rates of the reference report of agent i on a generic task k, while we use Z to denote this reference report, omitting the subscripts i and k, and use e_z^+, e_z^- to denote its error rates.

Our estimation algorithm resembles the "method of moments." We establish three equations on the first- to the third-order statistics, of which the parameters can be expressed by the unknown error rates e_z^+, e_z^-. We show that the three equations, with knowing the true parameters (which is true in the asymptotic setting), together uniquely determine e_z^+, e_z^-. Thus, we can solve the three equations to obtain e_z^+, e_z^-. In the next section, we argue that in the finite sample setting, with imperfect estimates of the parameters of the three questions, the solution from these three perturbed equations still approximate the true values of e_z^+, e_z^- with guaranteed accuracy.

To construct these three equations, we make the following preparation. Let $s_{j,k}$ be the realization of the prediction signal $S_{j,k}$ of agent j on task k, and let $\underline{\mathcal{S}_{-i} := \{s_{j,k}\}_{j\neq i, k\in[M]}}$ be the realization profile of all prediction signals from all peer agents of agent i. On a generic task k, we draw three random variables Z_1, Z_2, Z_3. Z_1 represents the realization of a prediction signal uniformly randomly drawn from the set of all prediction signals $\{s_{j,k}\}_{j\neq i}$ on task k except agent i's. Z_2 represents the realization of another uniformly randomly picked prediction signal from set $\{s_{j,k}\}_{j\neq i}$ but excluding Z_1. Similarly, Z_3 represents the realization of another uniformly randomly picked prediction signal from set $\{s_{j,k}\}_{j\neq i}$ but excluding Z_1 and Z_2. Because agents' reports are conditionally independent, Z_1, Z_2, Z_3 are also independent conditioned on the ground truth. Moreover, Z_1 and the reference report Z have the same error rates, as they are generated by the same random process. With infinite number of agents, Z_2 and Z_3 also have the same error rates as Z. Furthermore, (Z_1, Z_2, Z_3) is i.i.d. across different tasks according to Proposition 4.1 and Assumption 4. Therefore, with infinite number of tasks (and, thus, infinite number of samples from the joint distribution Z_1, Z_2, Z_3), we can know the exact distribution parameters of any statistics about Z_1, Z_2, and Z_3. We can then establish the following three equations.

---

(1) First-order equation: The first equation is based on the distribution of Z. Let $\underline{\alpha_{-i} := \Pr[Z = 1]}$. α_{-i} can be expressed as a function of e_z^+, e_z^- via spelling out the conditional expectation:

$$\alpha_{-i} = p \cdot \Pr[Z = 1|Y = 1] + (1 - p) \cdot \Pr[Z = 1|Y = 0] = p \cdot (1 - e_Z^+) + (1 - p) \cdot e_Z^-. \quad (4)$$

(2) Matching between two prediction signals: The second equation is based on a second-order statistic called the matching probability. We consider the matching-on-1 probability of Z_1, Z_2, that is, the matching-on-1 probability of the prediction signals from two uniformly randomly picked peer agents of agent i). Let $\underline{\beta_{-i} := \Pr[Z_1 = 1, Z_2 = 1]}$. It can be written as a function of e_z^-, e_z^+ as follows:

$$\begin{aligned}\beta_{-i} &= p \cdot \Pr\left[Z_1 = 1, Z_2 = 1|Y = 1\right] + (1 - p) \cdot \Pr\left[Z_1 = 1, Z_2 = 1|Y = 0\right] \\ &= p \cdot \Pr\left[Z_1 = 1|Y = 1\right] \cdot \Pr\left[Z_2 = 1|Y = 1\right] + (1 - p) \cdot \Pr\left[Z_1 = 1|Y = 0\right] \Pr\left[Z_2 = 1|Y = 0\right] \\ &= p \cdot (1 - e_z^+)^2 + (1 - p) \cdot (e_z^-)^2. \end{aligned} \quad (5)$$

(3) Matching among three prediction signals: The third equation is obtained by going one order higher. We check the matching-on-1 probability over three prediction signals Z_1, Z_2, Z_3 uniformly randomly drawn from three different peer agents on the same task. Let $\underline{\gamma_{-i} := \Pr[Z_1 = Z_2 = Z_3 = 1]}$. Similar to Equation (5), we have that

$$\gamma_{-i} = p \cdot (1 - e_z^+)^3 + (1 - p) \cdot (e_z^-)^3. \quad (6)$$

ALGORITHM 1: e_z^+, e_z^- solver

Input: $\alpha_{-i}, \beta_{-i}, \gamma_{-i}, \mathbb{1}(p > 0.5)$
Output: e_z^+, e_z^-

1: Compute the following quantities:

$$a := \frac{\gamma_{-i} - \alpha_{-i}\beta_{-i}}{\beta_{-i} - (\alpha_{-i})^2},\ b := \frac{\alpha_{-i}\gamma_{-i} - (\beta_{-i})^2}{\beta_{-i} - (\alpha_{-i})^2}.$$

2: Let

$$\underline{x} := \frac{a - \sqrt{a^2 - 4b}}{2},\ \overline{x} := \frac{a + \sqrt{a^2 - 4b}}{2}, p' := \frac{\alpha_{-i} - \underline{x}}{\overline{x} - \underline{x}}$$

3: If $\mathbb{1}(p' > 0.5) = \mathbb{1}(p > 0.5)$, then $e_z^+ = 1 - \overline{x},\ e_z^- = \underline{x}$, else $e_z^+ = 1 - \underline{x},\ e_z^- = \overline{x}$.

Note that all three parameters $\alpha_{-i}, \beta_{-i}, \gamma_{-i}$ can be perfectly estimated using $\mathcal{S}_{-i}$ with an infinite number of tasks and agents yet without accessing any of the ground truth. With the knowledge of these three parameters, we prove the following:

Theorem 6.3. *p, e_z^-, e_z^+ are uniquely identified by Equations (4) to (6) under Assumption 3 ($p \neq 0.5$ and the principal knows whether $p > 0.5$ or not). The solution is in the closed form shown in Algorithm 1.*

Proof. Let $x^- := e_z^-,\ x^+ := 1 - e_z^+$. Recall the three equations we have:

$$\alpha_{-i} = (1-p) \cdot x^- + p \cdot x^+ \tag{7}$$

$$\beta_{-i} = (1-p) \cdot (x^-)^2 + p \cdot (x^+)^2 \tag{8}$$

$$\gamma_{-i} = (1-p) \cdot (x^-)^3 + p \cdot (x^+)^3. \tag{9}$$

We can rewrite the three equations as

$$\alpha_{-i} - x^+ = (1-p)(x^- - x^+) \tag{10}$$

$$\beta_{-i} = (1-p)(x^- - x^+)(x^- + x^+) + (x^+)^2 \tag{11}$$

$$\gamma_{-i} = (1-p)(x^- - x^+)\left((x^-)^2 + x^- \cdot x^+ + (x^+)^2\right) + (x^+)^3. \tag{12}$$

Plugging Equation (10) into Equations (11) and (12) and reorganizing the two equations, we have, respectively, that

$$\beta_{-i} = \alpha_{-i}(x^- + x^+) - x^- \cdot x^+ \tag{13}$$

$$\gamma_{-i} = \alpha_{-i}\left((x^- + x^+)^2 - x^- \cdot x^+\right) - x^- \cdot x^+(x^- + x^+). \tag{14}$$

Let

$$x^- + x^+ = a,\ x^- \cdot x^+ = b.$$

Then, we have that $a = \frac{b+\beta_{-i}}{\alpha_{-i}}$ from Equation (13). Note that a is well defined, as o.w. if $\alpha_{-i} = 0$, we have to have that $x^- = x^+ = 0$, which leads to $e_z^- + e_z^+ = 1$, a contradiction.

Substituting $x^- + x^+$ and $x^- \cdot x^+$ with $\frac{b+\beta_{-i}}{\alpha_{-i}}$ and b correspondingly in Equation (14), we have that

$$\alpha_{-i} \cdot \left(\frac{(b+\beta_{-i})^2}{(\alpha_{-i})^2} - b\right) - b \cdot \frac{b+\beta_{-i}}{\alpha_{-i}} = \gamma_{-i} \tag{15}$$

$$\Rightarrow \frac{(b+\beta_{-i})^2}{\alpha_{-i}} - b\cdot\alpha_{-i} - \frac{b^2}{\alpha_{-i}} - \frac{b\cdot\beta_{-i}}{\alpha_{-i}} = \gamma_{-i} \tag{16}$$

$$\Rightarrow \left(\frac{\beta_{-i}}{\alpha_{-i}} - \alpha_{-i}\right) b = \gamma_{-i} - \frac{(\beta_{-i})^2}{\alpha_{-i}} \Rightarrow b = \frac{\alpha_{-i}\gamma_{-i} - (\beta_{-i})^2}{\beta_{-i} - (\alpha_{-i})^2}. \tag{17}$$

Thus, $a = \frac{b+\beta_{-i}}{\alpha_{-i}} = \frac{\gamma_{-i}-\alpha_{-i}\beta_{-i}}{\beta_{-i}-(\alpha_{-i})^2}, b = \frac{\alpha_{-i}\gamma_{-i}-(\beta_{-i})^2}{\beta_{-i}-(\alpha_{-i})^2}$. Then, from $x^- + x^+ = a,\ x^-\cdot x^+ = b$, we have that

$$x^+ = \frac{a\pm\sqrt{a^2-4b}}{2}, x^- = \frac{a\mp\sqrt{a^2-4b}}{2}, p = \frac{\alpha_{-i}-x^-}{x^+-x^-}.$$

Thus, we have two pairs of solutions for the error rates and the prior:

$$e^+_{z,(1)} = 1 - \frac{a+\sqrt{a^2-4b}}{2}, e^-_{z,(1)} = \frac{a-\sqrt{a^2-4b}}{2}, p_{(1)} = \frac{\alpha_{-i}-e^-_{z,(1)}}{1-e^+_{z,(1)}-e^-_{z,(1)}}$$

$$e^-_{z,(2)} = 1-e^+_{z,(1)}, e^+_{z,(2)} = 1-e^-_{z,(1)}, p_{(2)} = 1-p_{(1)}.$$

As in these two solutions, the values for the prior is symmetric with regard to 0.5. Thus, by Assumption 3, the principal can identify the unique correct solution from the two. □

We can continue to establish higher-order equations. However, we show that they do not provide additional information about the three unknown variables, p, e^+_z, and e^-_z.

THEOREM 6.4. *Any higher-order (≥4) matching equations can be expressed by the first- to the third-order equations, Equations (4) to (6).*

PROOF. We follow the shorthand notations as in the proof of Theorem 6.3. The n-th equation is

$$\Pr[Z_1 = \cdots = Z_n = 1] = (1-p)(x^-)^n + p(x^+)^n.$$

For $n \geq 4$, the right-hand side of the equation can be expressed as

$$\begin{aligned}(1-p)(x^-)^n + p(x^+)^n &= \left((1-p)(x^-)^{n-1} + p(x^+)^{n-1}\right)(x^- + x^+)\\ &\quad - x^-\cdot x^+\left((1-p)(x^-)^{n-2} + p(x^+)^{n-2}\right)\\ &= \Pr[Z_1 = \cdots = Z_{n-1}](x^- + x^+)\\ &\quad - \Pr[Z_1 = \cdots = Z_{n-2}]x^-\cdot x^+.\end{aligned}$$

As we know from the proof of Theorem 6.3, $x^- + x^+$ and $x^-\cdot x^+$ are uniquely determined by the first three equations, ithat is, Equations (4) to (6) (no matter whether Assumption 3 is made or not). Therefore, by induction starting from $n = 4$, the n-th equation can be expressed by the first three equations. □

Now we have completed our SSR mechanisms. The full version of the mechanisms is presented in Mechanism 2. Intuitively speaking, Theorem 6.3 shows that without ground truth data, knowing how frequently agents' predictions reach consensus with each other will help us characterize the (average) subjective biases in their reports. Furthermore, it implies that SSR mechanisms are asymptotically (in m, n) preserving the information quantification property that SPSRs have, that is, $\mathbb{E}_Z[R(q_{i,k}, Z)] = \mathbb{E}_Y[S(q_{i,k}, Y)]$, and that SSR mechanisms induce truthful reporting as the unique best uniform strategy for an agent, when Z is informative (i.e., $1 - e^+_z - e^-_z \neq 0$), and as a best strategy otherwise. Formally, we have the following theorem.

Mechanism 2: SSR mechanisms

1: For each task k, uniformly randomly pick at least 3 agents, assign task k to them, collect their reported predictions $q_{i,k}$, and generate the prediction signal $S_{i,k}$ for each prediction.
2: For each agent i and each task k the agent answers, uniformly randomly select one prediction signal $S_{j,k}$ from the agent's peers' prediction signals on the same task and let the reference report $Z_{i,k} := S_{j,k}$.
3: Establish Equations (4) to (6) and solve out the error rates $e^-_{z_i}, e^+_{z_i}$ for $Z_{i,k}$ for any k using Algorithm 3.
4: Pay each agent i's prediction $q_{i,k}$ on each task k the agent answers by applying SSR_α with $q_{i,k}$ and the noisy estimate $Z_{i,k}$ with error rates $e^+_{z_i}, e^-_{z_i}$ if $e^+_{z_i} + e^+_{z_i} \neq 1$, and pay 0, otherwise.

Theorem 6.5. *Under Assumptions 1 to 4, SSR mechanisms are dominant uniform strategy truthful with an infinite number of tasks and agents. Furthermore, for any agent i and task k, if the average prediction of all other agents is informative, that is, $e^+_z + e^-_z \neq 1$ for the noisy estimate of the ground truth $Z_{i,k}$ constructed for agent i, then the expected score of SSR mechanisms for agent i's prediction on a task is equal to the expected score given by the corresponding SPSR S: $\forall q_{i,k} \in [0,1], \mathbb{E}_{Z_{i,k}}[R(q_{i,k}, Z_{i,k})] = \mathbb{E}_{Y_k}[S(q_{i,k}, Y_k)]$.*

Proof. Recall that in Assumption 4, we assume that each agent adopts the same reporting strategy across tasks. As long as this assumption is satisfied, for an agent i, no matter what exact strategy the other agents play, we can always correctly estimate the error rates e^+_z and e^-_z of the reference report Z constructed for agent i, according to Theorem 6.3. Furthermore, by Lemma 6.1, Z is independent from agent i's belief conditioned on the ground truth. Therefore, according to Theorem 5.6, when $e^+_z + e^-_z \neq 1$, that is, the other agents' average prediction is informative about the ground truth Y, SSRs give agent i's prediction $q_{i,k}$ a reward unbiased to the expected reward given by the corresponding SPSR, that is, $\forall q_{i,k}, \mathbb{E}_{Z_{i,k}}[R(q_{i,k}, Z_{i,k})] = \mathbb{E}_{Y_k}[S(q_{i,k}, Y_k)]$. Consequently, truthful reporting strictly maximizes the expected reward of agent i. When $e^+_z + e^-_z = 1$, that is, the other agents' average prediction is uninformative about Y_k for task k, SSR mechanisms always reward agent i zero, where truthful reporting also maximizes the expected reward of agent i. Thus, SSR mechanisms are dominant uniform strategy truthful. □

Remark 1. Theorems 6.3 and 6.5 rely on Proposition 4.1 and Assumptions 3 and 4. Proposition 4.1 and Assumption 4 guarantee that there is a similar information pattern across the predictions of different tasks that we can learn to infer the ground truth. Therefore, they can be hardly relaxed in IEVW settings. For Assumption 3, we'd like to argue that at least one bit of information is needed in order to distinguish the case in which agents are truthfully reporting from the case in which agents are misreporting by reverting their observations. This is because for any distribution of the observed reports of agents resulting from a world with parameters (p, e^+_z, e^-_z) and with agents reporting truthfully, there always exists the following counterfactual world achieving the same distribution of the observed reports of agents: a world with parameters $(1-p, 1-e^-_z, 1-e^+_z)$ and with agents misreporting predictions via relabelling $0 \to 1$ and $1 \to 0$. Thus, the mechanism designer cannot tell the two worlds apart from only the observed reports. Some studies [20, 23] relax Assumption 3 by allowing the truthful reporting strategy to weakly dominate this "relabeling equilibrium."

We will show in the next section that SSR mechanisms are also dominant uniform strategy truthful with a finite number of tasks and agents under mild conditions. Several remarks follow. (1) We would like to emphasize again that for an agent i, both Z and $R(\cdot)$ come from the prediction signals of the agent's peer agents' reports $\mathcal{S}_{-i}$: Z is directly picked from $\mathcal{S}_{-i}$; $R(\cdot)$ depends on the error rates e_z^+ and e_z^- of Z, which are also learnt from $\mathcal{S}_{-i}$. (2) When making reporting decisions under SSR mechanisms, agents can choose to be oblivious of how much error presents in others' reports, because truthful reporting is the dominant strategy, that is, no matter what uniform reporting strategy other agents play, truthful reporting always maximizes the expected reward. This removes the practical concern of implementing truthful reporting as a particular Nash Equilibrium when there exists a non-truthful reporting equilibrium. (3) Another salient feature of SSR mechanisms is that they transfer the cognitive load of having prior knowledge from the agent side to the mechanism designer side. Yet we do not assume that the designer has exact knowledge of the prior either (but the knowledge of whether the prior is greater than 0.5 or not). Instead, we will leverage the power of estimation from reported data to achieve our goals.

6.3 Finite Sample Analysis

With finite m, n, we use the same procedure as shown in Algorithm 1 to estimate the error rates e_z^+, e_z^- for each agent, except that we cannot have the exact value for $\alpha_{-i}, \beta_{-i}, \gamma_{-i}$ but only with finite-sample estimates for them. Specifically, for agent i, letting k_1, k_2, k_3 (which could be different on different tasks) be the three agents whose prediction signals are selected as Z_1, Z_2, Z_3 on each task $k \in [M]$,[3] we estimate that

$$\widetilde{\alpha_{-i}} = \frac{\sum_{k=1}^m \mathbb{1}(S_{k_1,k} = 1)}{m}, \ \widetilde{\beta_{-i}} = \frac{\sum_{k=1}^m \mathbb{1}(S_{k_1,k} = S_{k_2,k} = 1)}{m}, \ \widetilde{\gamma_{-i}} = \frac{\sum_{k=1}^m \mathbb{1}(S_{k_1,k} = S_{k_2,k} = S_{k_3,k} = 1)}{m}.$$

We then use these three values to replace $\alpha_{-i}, \beta_{-i}, \gamma_{-i}$, respectively, in Algorithm 1 to solve Equations (4) to (6). We denote the resulting error rates as $\widetilde{e_z^+}$ and $\widetilde{e_z^-}$ and the corresponding SSR_α using these error rates as $\widetilde{R}(\cdot)$.

There are two reasons that these finite-sample estimates $\widetilde{e_z^+}$ and $\widetilde{e_z^-}$ are not equal to the exact true error rates e_z^+ and e_z^- for Z. First, in constructing Equations (4) to (6), the error rates of two randomly picked prediction signals Z_2, Z_3 will not have the exact same error rates with Z, as these signals come from a slightly different agent pool. Second, $\widetilde{\alpha_{-i}}, \widetilde{\beta_{-i}}, \widetilde{\gamma_{-i}}$ are not exactly equal to $\alpha_{-i}, \beta_{-i}, \gamma_{-i}$ with finite samples. However, we will show that the errors induced by these two factors in estimating the error rates diminish with m and n. Consequently, the SSR computed using $\widetilde{e_z^+}, \widetilde{e_z^-}$ also have a small and diminishing error towards the SSR computed with the exact error rates e_z^+, e_z^-.

Lemma 6.6. *$\widetilde{e_z^+}, \widetilde{e_z^-}$ given by Algorithm 1 using $\widetilde{\alpha_{-i}}, \widetilde{\beta_{-i}}, \widetilde{\gamma_{-i}}$ satisfy that for an arbitrary $\delta \in (0,1)$, with probability at least $1-\delta$, $|\widetilde{e_z^+} - e_z^+| \le \epsilon$, $|\widetilde{e_z^-} - e_z^-| \le \epsilon$ for some $\underline{\epsilon = O(\frac{1}{n} + \sqrt{\frac{\ln \frac{1}{\delta}}{m}})}$, which can be made arbitrarily small with increasing m and n.*

Proof Sketch. We present the high-level idea of our proof here and defer the complete proof to the appendix. We consider the two aforementioned errors separately. Both can be transformed to a diminishing error attaching to the evaluation of α_{-i}, β_{-i}, and γ_{-i}. This diminishing noise in α_{-i}, β_{-i}, and γ_{-i} can then be transformed into a diminishing error in the final solution of e_z^+, e_z^-. □

Next, we show that the deviations of the rewards of SSR mechanisms due to the imperfect estimation of the error rates in the finite sample case can also be bounded to be arbitrarily small. We first deal with a special case: even if $1 - e_z^+ - e_z^-$ is far from zero, the estimated $1 - \widetilde{e_z^+} - \widetilde{e_z^-}$ in the denominator of SSR_α can be arbitrarily close to zero by coincidence. In this case, agents can have

[3]In practice, we need to assign task k to these three randomly selected agents only.

unbounded scores, which may be far from the exact scores that agents should obtain when the estimation is perfect. To address this special case, the principal can select a threshold κ greater than but close to zero and pay agents zero when $|1-\widetilde{e_z^+}-\widetilde{e_z^-}|<\kappa$ instead of just when $1-\widetilde{e_z^+}-\widetilde{e_z^-}=0$. As a result, the final reward of each agent is always bounded. Next, we introduce a lemma that we will use in our proof.

Lemma 6.7. $\forall l_1, l_2, t_1, t_2 \in [-1,1], t_1, t_2 \neq 0, \left|\frac{l_1}{t_1}-\frac{l_2}{t_2}\right| \leq \frac{|l_1-l_2|+|t_1-t_2|}{|t_1 t_2|}$

Proof. $\left|\frac{l_1}{t_1}-\frac{l_2}{t_2}\right| = \left|\frac{l_1t_2-l_2t_1}{t_1t_2}\right| = \left|\frac{l_1t_2-l_2t_2+l_2t_2-l_2t_1}{t_1t_2}\right| \leq \frac{|t_2||l_1-l_2|+|l_2||t_2-t_1|}{|t_1t_2|} \leq \frac{|l_1-l_2|+|t_1-t_2|}{|t_1t_2|}$ □

This lemma is an extension of Lemma 7 in [25], which considers the case in which all variables are non-negative. Now we present our main theorem about the diminishing error in estimating the SSR scores.

Theorem 6.8. *For a bounded SPSR $S(\cdot)$ with supremum $\max S$, for an arbitrary $\delta \in (0,1)$, and some $\epsilon = O(\frac{1}{n}+\sqrt{\frac{\ln\frac{1}{\delta}}{m}})$ such that with probability at least $1-\delta$, $|\widetilde{e_z^+}-e_z^+|\leq\epsilon$, $|\widetilde{e_z^-}-e_z^-|\leq\epsilon$, let m and n be sufficiently large such that $\epsilon \leq |1-e_z^- - e_z^+|/4$, the SSR mechanism built upon $S(\cdot)$ satisfies, with probability at least $1-\delta$, that*

$$|\widetilde{R}(q_{i,k},Z)-R(q_{i,k},Z)| \leq \frac{12\max S}{\Delta^2}\cdot\epsilon,\ \forall i\in[n], k\in[m], q_{i,k}\in[0,1], Z\in\{0,1\},$$

where $\Delta = |1-e_z^- - e_z^+|$. Furthermore, taking over all of the randomness in the score, we have that

$$\left|\mathbb{E}[\widetilde{R}(q_{i,k},Z)]-\mathbb{E}[S(q_{i,k},Z)]\right| = O\left(\frac{1}{N}+\sqrt{\frac{\ln m}{m}}\right), \forall i,k.$$

Proof. This proof is straightforward following the error rate bounding result (Lemma 6.6). We use $sgn(Z), Z\in\{0,1\}$ as the superscript, where $sgn(0)$ refers to superscript "−" and $sgn(1)$ refers to superscript "+".

Consider an arbitrary agent i and a task k; we have that

$$\begin{aligned}|\widetilde{R}(q_{i,k},Z)-R(q_{i,k},Z)| &= \left|\left(\frac{1-\widetilde{e_z^{sgn(1-Z)}}}{1-\widetilde{e_z^+}-\widetilde{e_z^-}}-\frac{1-e_z^{sgn(1-Z)}}{1-e_z^+-e_z^-}\right)S(q_{i,k},Z)\right.\\ &\quad \left.-\left(\frac{\widetilde{e_z^{sgn(Z)}}}{1-\widetilde{e_z^+}-\widetilde{e_z^-}}-\frac{e_z^{sgn(Z)}}{1-e_z^+-e_z^-}\right)S(q_{i,k},1-Z)\right|\\ &\leq \left|\frac{1-\widetilde{e_z^{sgn(1-Z)}}}{1-\widetilde{e_z^+}-\widetilde{e_z^-}}-\frac{1-e_z^{sgn(1-Z)}}{1-e_z^+-e_z^-}\right|\max S\\ &\quad +\left|\frac{\widetilde{e_z^{sgn(Z)}}}{1-\widetilde{e_z^+}-\widetilde{e_z^-}}-\frac{e_z^{sgn(Z)}}{1-e_z^+-e_z^-}\right|\max S.\end{aligned}$$

Since $\epsilon \leq (1-e_z^- - e_z^+)/4$, we know that

$$|1-\widetilde{e_z^+}-\widetilde{e_z^-}| \geq |1-e_z^- - e_z^+|/2.$$

Thus, with probability at least $1-\delta$,

$$\left|\frac{1-\widetilde{e_z^{sgn(1-Z)}}}{1-\widetilde{e_z^+}-\widetilde{e_z^-}}-\frac{1-e_z^{sgn(1-Z)}}{1-e_z^+-e_z^-}\right| \le \frac{\left|\widetilde{e_z^{sgn(1-Z)}}-e_z^{sgn(1-Z)}\right|+\left|\widetilde{e_z^+}+\widetilde{e_z^-}-e_z^+-e_z^-\right|}{|(1-\widetilde{e_z^+}-\widetilde{e_z^-})(1-e_z^+-e_z^-)|}$$
$$\le \frac{3\epsilon}{|(1-\widetilde{e_z^+}-\widetilde{e_z^-})(1-e_z^+-e_z^-)|} \le \frac{6\epsilon}{\Delta^2}.$$

In these inequalities, the first "≤" follows Lemma 6.7, the second follows Lemma 6.6, and the third follows $|1-\widetilde{e_z^+}-\widetilde{e_z^-}| \ge |1-e_z^- - e_z^+|/2$. Similarly, we have that

$$\left|\frac{\widetilde{e_z^{sgn(Z)}}}{1-\widetilde{e_z^+}-\widetilde{e_z^-}}-\frac{e_z^{sgn(Z)}}{1-e_z^+-e_z^-}\right| = \frac{6\epsilon}{\Delta^2}.$$

Plugging back, we have proved the claim that with probability at least $1-\delta$,

$$|\widetilde{R}(q_{i,k},Z)-R(q_{i,k},Z)| \le \frac{12\epsilon\cdot\max S}{\Delta^2},\ \forall q_{i,k}\in[0,1], Z\in\{0,1\}.$$

As $\mathbb{E}[S(q_{i,k},Z)] = \mathbb{E}[R(q_{i,k},Z)]$, letting $\delta = \frac{1}{m}$, we have the expected error $|\mathbb{E}[\widetilde{R}(q_{i,k},Z)] - \mathbb{E}[S(q_{i,k},Z)]|$ bounded by $O((1-\frac{1}{m})(\frac{1}{n}+\sqrt{\frac{\ln m}{m}})+\frac{1}{m}) = O(\frac{1}{m}+\sqrt{\frac{\ln m}{m}})$. □

Theorem 6.8 indicates that the errors of the expected scores given by SSR mechanisms wwith regard to the expected score given by the underlying SPSR can be made arbitrarily small with sufficiently large m and n. As a result, for arbitrarily discretized report space of a prediction, SSR mechanisms are still dominant uniform strategy truthful with finite but sufficiently large m and n. To see this, we can make the error smaller than the minimum absolute difference of the SPSRs of any two allowed probability reports. In this way, there will be no beneficial deviation for agents to report non-truthfully. This result considers the reality that in real surveys, agents are often allowed to specify at most two decimal digits for probabilistic predictions.

Corollary 6.9. *For discretized report space of probabilistic predictions, SSR mechanisms that are built upon bounded SPSRs are dominant uniform strategy truthful for finite but sufficiently large m and n.*

7 GENERALIZATIONS TO MULTI-OUTCOME TASKS

In this section, we discuss how to extend SSR and SSR mechanisms to the multi-outcome, multi-task setting. A multi-outcome task asks agents to provide predictions about a multi-outcome random variable Y, which takes value from a finite support set $[c] = \{0,\dots,c-1\}$ with $c > 2$. A noisy estimate $Z \in [c]$ of the ground truth Y is characterized by a confusing matrix:

$$E_z = \begin{bmatrix} e_{0,0} & e_{0,1} & \cdots & e_{0,c-1} \\ e_{1,0} & e_{1,1} & \cdots & e_{1,c-1} \\ \cdots & \cdots & \cdots & \cdots \\ e_{c-1,0} & e_{c-1,1} & \cdots & e_{c-1,c-1} \end{bmatrix},$$

where $e_{u,v}$ represents the flipping probability of Z with regard to Y, that is, $e_{u,v} = \Pr[Z=v|Y=u], \forall u,v\in[c]$.

7.1 Generalization of SSR

The SSRs for a task with c outcomes are defined as follows. Let Δ^{c-1} be the $(c-1)$-dimension probability simplex, that is, $\Delta^{c-1} := \left\{(x_0, \ldots, x_{c-1}) | \sum_{i=0}^{c-1} x_i = 1, x_0, \ldots, x_{c-1} \geq 0\right\}$.

Definition 7.1 (Surrogate Scoring Rules). $R : \Delta^{c-1} \times [c] \to \mathbb{R}$ is an SSR for a c-outcome task if for some strictly proper scoring rule $S : \Delta^{c-1} \times [c] \to \mathbb{R}$ and a strictly increasing function $f : \mathbb{R} \to \mathbb{R}$, the following equation holds:

$$\forall \mathbf{p}, \mathbf{q} \in \Delta^{c-1}, \forall E_z \in [0,1]^{c \times c} (E_z \text{ is invertible}) : \underline{\mathbb{E}_Z[R(\mathbf{q}, Z)] = f\left(\mathbb{E}_Y[S(\mathbf{q}, Y)]\right)},$$

where the ground truth Y is drawn from Categorical($\mathbf{p}$) and Z is a noisy estimate of Y with confusing matrix E_z.

We have the following theorem immediately.

Theorem 7.2. *Given the prior* $\mathbf{p}$ *of the ground truth* Y *and a private signal* O_i*, SSR* $R(\mathbf{q}, z)$ *with a noisy estimate* Z *of the ground truth is strictly proper for eliciting an agent's posterior* $\mathbf{p}_i := \Pr[Y|O_i]$ *if* Z *and* O_i *are independent conditioned on* Y *and* E_z *is invertible.*

Now we give an implementation of SSR, SSR_α, for a c-outcome task. Let $S(\mathbf{q}_i)$ be the vector of SPSR scores for a prediction $\mathbf{q}_i \in \Delta^{c-1}$ under each realization of Y, that is, $S(\mathbf{q}_i) := (S(\mathbf{q}_i, Y = 0), \ldots, S(\mathbf{q}_i, Y = c-1))$. Similarly, let $R(\mathbf{q}_i) := (R(\mathbf{q}_i, Z = 0), \ldots, R(\mathbf{q}_i, Z = c-1))$. Our implementation SSR_α goes as follows:

$$R(\mathbf{q}_i) := (E_z)^{-1} S(\mathbf{q}_i).$$

Clearly, for SSR_α we have that $S(\mathbf{q}_i) = E_z \cdot R(\mathbf{q}_i)$, which gives

$$\forall v \in [c], S(\mathbf{q}_i, Y = v) = \sum_{k=0}^{c-1} e_{v,k} R(\mathbf{q}_i, Z = k) = \mathbb{E}_{Z|Y=v}[R(\mathbf{q}_i, Z)].$$

Lemma 7.3. *For* SSR_α*:* $\forall v \in [c], \mathbb{E}_{Z|Y=v}[R(p_i, Z)] = S(p_i, Y = v)$.

Theorem 7 follows immediately.

Theorem 7.4. SSR_α *is an SSR for a multi-outcome task, and for any distribution* $\mathbf{p} \in \Delta^{c-1}$ *of the ground truth* Y *and for any invertible confusing matrix* E_z *of a noisy estimate* Z *of the ground truth, we have that* $\forall \mathbf{q} \in \Delta^{c-1}, \mathbb{E}_Z[R(\mathbf{q}, Z)] = \mathbb{E}_Y[S(\mathbf{q}, Y)]$.

A detailed example of SSR_α for a three-outcome task follows.

Example 7.5. Let $c = 3$ and let the confusing matrix of a noisy signal Z be

$$E_z = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix} \Rightarrow (E_z)^{-1} = \begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}.$$

We obtain a closed form of SSR_α:

$$\begin{aligned} R(\mathbf{q}, Z = 0) &:= 3S(\mathbf{q}, 0) - S(\mathbf{q}, 1) - S(\mathbf{q}, 2) \\ R(\mathbf{q}, Z = 1) &:= -S(\mathbf{q}, 0) + 3S(\mathbf{q}, 1) - S(\mathbf{q}, 2) \\ R(\mathbf{q}, Z = 2) &:= -S(\mathbf{q}, 0) - S(\mathbf{q}, 1) + 3S(\mathbf{q}, 2) \end{aligned}$$

7.2 Generalization of SSR Mechanisms

SSR mechanisms can also be extended to multi-outcome tasks and maintain the two properties we pursue: the dominant uniform strategy truthfulness and qualifying the value of information as what SPSRs do.

We consider the same setting of information structures under Assumptions 1 to 3, except that $Y_k, k \in [m]$ in these assumptions are c-outcome categorical random variables, agents' beliefs are categorical distributions, and, in Assumption 3, the prior probabilities of Y_k being each outcome is different and the principal knows the order of these prior probabilities. As we have shown that SSRs can be extended to multi-outcome events, to construct the corresponding SSR mechanism, we just need to construct the corresponding noisy estimate Z of the ground truth and estimate the confusion matrix E_z for multi-outcome tasks.

The noisy estimate Z for an agent i on task k can be constructed similar to that of the counterpart in the binary case, that is, we uniformly randomly pick an agent $j \neq i$ and draw $Z \sim \text{Categorical}(\mathbf{q}_{j,k})$, where $\mathbf{q}_{j,k}$ is the reported distribution of Y_k from agent j. Then, the confusion matrix can also be estimated using the method of moments. However, as there are $c^2 - 1$ unknown parameters in the confusion matrix E_z and the prior $\mathbf{p}$ of Y_k, we have to establish $c^2 - 1$ equations. These equations could be solved numerically. These $c^2 - 1$ equations will have $c!$ real-value symmetric solutions, each corresponding to a permutation of the labeling of the c outcomes. To identify the unique solution that yields the true confusion matrix and the prior of Y, that is, to identify the correct labeling of the outcomes, the principal has to know the order of the prior probabilities of Y_k being each outcome, as what we assume in Assumption 3 for multi-outcome tasks. Thus, with the multi-task variant of Assumptions 1 to 3, we can still construct a noisy estimate Z of the ground truth, estimate its confusion matrix, and apply SSRs to obtain unbiased estimates of agents' scores given by the underlying SPSRs.

Despite the positive result in theory, there are some caveats of applying SSR mechanisms to multi-outcome tasks. First, Assumption 1 essentially assumes that the confusion matrix of an agent is homogeneous across different tasks. However, as there is no clear correspondence between the labels of the outcomes of different tasks, the confusion matrix of the noisy estimate Z for an agent is less likely to be homogeneous across different tasks. Therefore, the real data can deviate far from Assumption 1. Second, as there are more parameters in the confusion matrix to estimate in the multi-task case than in the binary case, we need a much larger number of agents and tasks and denser predictions to maintain decent estimation accuracy. Third, to apply an SSR mechanism to multi-outcome tasks, these tasks have to have the same number of outcomes. However, in most crowd forecasting projects, the number of multi-outcome tasks with the same number of outcomes is much smaller than the number of binary questions and may not be sufficient to make an accurate estimation of the confusion matrix. These caveats leave a massive space for future research.

8 EMPIRICAL STUDIES

Using 14 real-world human forecasting datasets, we empirically examine the performance of SSR mechanisms in revealing agents' prediction accuracy in terms of SPSRs. We focus on three aspects: the unbiasedness of SSR, the correlation of SSR scores to SPSR scores, and the accuracy of SSR in selecting true top forecasters in terms of SPSRs. We also compare the performance of SSR mechanisms to several existing peer prediction mechanisms. The overall results show that our SSR mechanisms have an advantage in recovering SPSRs.

8.1 Setting

8.1.1 Datasets. We conduct our experiments on 14 datasets from three human forecasting and crowdsourcing projects: the Good Judgment Project (GJP), the Hybrid Forecasting Competition

(HFC), and the human judgment datasets collected by the Massachusetts Institute of Technology (MIT). These three projects differ in participant population, forecasting topics, and elicitation methods, offering a rich environment for empirical evaluation.

GJP datasets [3]. The GJP data consists of four datasets for geopolitical forecasting questions. The four datasets, denoted by G1~G4, were collected from 2011 to 2014, respectively. They contain different sets of forecasting questions and forecasters.

HFC datasets [16]. We use the forecast data of team participants in the Hybrid Forecasting Competition. The data consist of three datasets, denoted by H1~H3, referring to the forecasting data collected in the preseason competition, the first competition, and the second competition, respectively. The the preseason competition lasted half a year, and the two formal competitions lasted around one year. The three datasets have different forecasting questions and partially overlapped participating teams.

MIT datasets [32]. The MIT data consist of seven datasets, denoted by M1a, M1b, M1c, M2, M3, M4a, M4b. Each dataset uses one of four sets of questions and has a different participant pool. The questions range from guessing the capital of each state and predicting the price interval of artworks to some trivia questions. The forecasters were students in class and colleagues in labs. In datasets M1a, M1b, M4a, M4b, forecasters report only binary votes on forecasting questions. In datasets M1c, M2, M3, forecasters give probabilistic predictions.

Both GJP and HFC allow participants to make daily forecasts. For testing peer prediction mechanisms in our setting, we need to use only a single prediction for each participant on a forecasting question. In our experiments, we mainly focus on the final prediction of each participant made on each question (i.e., the last prediction made by each participant before the close date of the corresponding forecasting question) in these two projects. At the end of Section 8.2, we complement our analysis by verifying the robustness of SSRs with respect to the choice of the time the predictions are made. Also, we focus on the forecasting questions that have binary outcomes in these datasets. To have a relatively stable estimation over the accuracy of agents, we filter out participants who made predictions on less than 15 questions. The basic statistics of these datasets are presented in Table 1.

8.1.2 SPSRs. We consider three SPSRs — the Brier score, the log scoring rule, and the rank-sum scoring rule –because of their usage in practice and connections to machine learning concepts. The first two are the most widely adopted scoring rules. They are equivalent to two main loss functions, the squared error and the cross-entropy loss, respectively, used in the machine learning community. The rank-sum scoring rule can be written as an affine transformation of the area under the receiver operating characteristic curve (AUC-ROC),[4] which is also a widely adopted accuracy metric in the machine learning community.

In our experiments, we adopt the conventional formula of the Brier score used in the GJP and HFC projects. The Brier score ranges from 0 to 2, with a smaller score corresponding to higher accuracy. This is different from using SPSRs as a payment method, in which case, the higher the better. We can transfer between these two usages by applying a negative scalar. We orient the log scoring rule and the rank-sum score rule in the same direction as the Brier score, with a minimum (best) score of 0. The exact formula for each scoring rule is as follows: Recall that $q_{i,k}$ and Y_k

[4]The affine transformation coefficients are determined by the numbers of tasks with ground truth 1 and ground truth 0 according to Equations (12) and (13) [30]. Thus, when evaluating agents' prediction accuracy on the same set of answered questions, the rank-sum scoring rule is equal to the AUC-ROC for each agent up to the same affine transformation determined by the ground truth of the questions. However, the AUC-ROC itself is not an SPSR, as when considering the incentive, the affine transformation coefficients may differ in different agents' beliefs.

Table 1. Statistics about Binary-Outcome Datasets from GJP, HFC, and MIT Datasets

Items	G1	G2	G3	G4	H1	H2	H3	M1a	M1b	M1c	M2	M3	M4a	M4b
# of questions (original)	94	111	122	94	44	86	203	50	50	50	80	80	90	90
# of agents (original)	1,972	1,238	1,565	7,019	79	317	222	51	32	33	39	25	20	20
After applying the filter														
# of questions	94	111	122	94	44	86	203	50	50	50	80	80	90	90
# of agents	1,409	948	1,033	3,086	79	316	222	51	32	33	39	25	20	20
Avg. # of answers per question	851	533	369	1,301	71	295	220	51	32	33	39	18	20	20
Avg. # of answers per agent	57	62	44	40	39	80	201	50	50	50	80	60	90	90
Majority vote correct ratio (%)	0.90	0.92	0.95	0.96	0.93	0.93	0.86	0.58	0.76	0.74	0.61	0.68	0.62	0.72

are agent i's prediction and the ground truth for task k, respectively, and $[m_i]$ is the set of tasks answered by agent i.

- **Brier score**: $S^{\text{Brier}}(q_{i,k}, Y_k) = 2(q_{i,k} - Y_k)^2$. We use the mean Brier score, $\frac{1}{m_i} \cdot \sum_{k \in [m_i]} S^{\text{Brier}}(q_{i,k}, Y_k)$, to represent an agent's overall accuracy under the Brier score over the set of tasks the agent answered.
- **Log scoring rule**: $S^{\log}(q_{i,k}, Y_k) = Y_k \log(q_{i,k}) + (1 - Y_k)\log(1 - q_{i,k})$. We use the mean log score, $\frac{1}{m_i} \sum_{k \in [m_i]} S^{\log}(q_{i,k}, Y_k)$, to represent an agent's overall accuracy under the long scoring rule over the tasks the agent answered. As the log scoring rule is unbounded when the forecast predicts the opposite of the ground truth, we change all forecasts of 1 to 0.99 and forecasts of 0 to 0.01 to ensure that the score is always a real number.
- **Rank-sum scoring rule** is a multi-task scoring rule. For a single task k, it assigns a score

$$S^{\text{rank}}(q_{i,k}, y_k) = -y_k \cdot \psi\left(q_{i,k} | \{q_{i,k'}\}_{k' \in [m_i]}\right),$$

  where $\psi(q_{i,k} | \{q_{i,k'}\}_{k' \in [m_i]}) := \sum_{k' \in [m_i]} \mathbb{1}(q_{i,k'} < q_{i,k}) - \sum_{k' \in [m_i]} \mathbb{1}(q_{i,k'} > q_{i,k})$ is the rank of prediction $q_{i,k}$ among all predictions from agent i. Then, agent i's rank-sum score S_i^{rank} is defined as $S_i^{\text{rank}} = \sum_{k \in [m_i]} S^{\text{rank}}(q_{i,k}, Y_k)$.[5] The range of the score increases with the number of answered tasks quadratically thus; we use the normalized score $1 + \frac{4}{m_i^2} S_i^{\text{rank}}$ with range $[0, 2]$.

8.1.3 Treatments. Though existing peer prediction methods are not designed for recovery of SPSRs, we add comparisons to them for completeness of our study.[6] In particular, we would like to understand whether in practice SSR mechanisms have the advantage of revealing the true scores given by SPSRs while not accessing ground truth information.

In our experiments, we consider four popular existing peer prediction methods serving as comparisons to SSRs: proxy scoring rules (PSRs) with extremized mean [47], peer truth serum (PTS) [35], correlated agreement (CA) [42], and determinant mutual information (DMI) [20].

PSRs are to directly apply the SPSRs with respect to an unbiased proxy of the ground truth. When the principal knows no unbiased proxy, Witkowski et al. [47] recommend using the extremized mean of the reported predictions to serve as the proxy. In our experiments, we adopt the same formula for the extremized mean as in their experiments [47], that is, $\frac{\bar{q}_k^2}{\bar{q}_k^2 + (1-\bar{q}_k)^2}$, where $\bar{q}_k$ is the average reported prediction on task k. Using different SPSRs as the underlying scoring rule, we can get different PSRs and SSRs.

[5] The AUC-ROC of agent i is $\frac{1}{2}(1 - \frac{1}{m_i^+(m_i - m_i^+)} S_i^{\text{rank}})$, where $m_i+ := \sum_{k' \in [m_i]} \mathbb{1}(Y_{k'} = 1)$ (given by Equations (12) and (13), [30]).

[6] We do not intend to claim that our mechanism is better in any sense, as it would be an unfair comparison since the goals were different in each design of these mechanisms. For example, the mechanisms [20, 42] can characterize determinant mutual information between an agent's reports and the underlying ground truth.

PTS, CA, and DMI do not depend on SPSRs and are designed to elicit categorical labels instead of probabilistic predictions. Thus, we make the following adaptation for them to take probabilistic predictions as inputs. Our adaptation is based on the fact that, in essence, these mechanisms all appreciate the joint distribution of agents' reported labels to compute the scores: For a task k, an agent who reports probability $P_{i,k}$ believes that the true label of the task has probability $P_{i,k}$ to be 1. Therefore, on this task, the joint probability of agent i's believed true label and agent j's believed true label both being 1 is $P_{i,k}P_{j,k}$, assuming their believed true labels are independent conditioned on their predictions. In this way, we can compute the joint distribution of the believed true labels of two peer agents on each task and their joint distribution over the whole dataset is the mean of their joint distributions on each task. Using this joint distribution over the whole dataset, we can compute the scores for PTS, CA, and DMI directly. This adaptation method for PTS, CA, and DMI turns out to give better correlations between the scores of these three mechanisms and the true SPSR scores than the alternative adaptation method of using the most likely categorical labels indicated by the probabilistic predictions as inputs for these mechanisms (see how the correlations shown in Figures A.1 and A.2 in the appendix – most likely labels as inputs – compare to the correlations shown in Figures 2 and 3).

8.2 Main Results

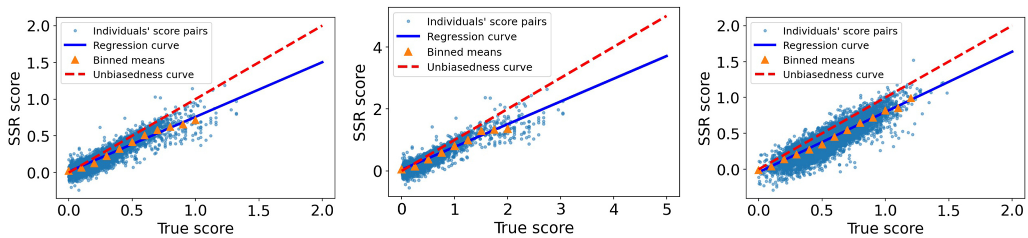
Unbiasedness of SSR. Our theorem shows that under certain assumptions, the reward of an SSR mechanism is unbiased to the reward of the SPSR that the SSR mechanism is built upon. However, it is unclear to what extent this unbiasedness holds in real datasets for which these assumptions are unlikely to hold strictly. Therefore, we empirically examine the concrete relationship between SSR scores and the corresponding SPSR scores.

Figure 1 plots the score pairs received by forecasters in each of the 14 datasets. Each score pair represents the SPSR score and the SSR score that an individual forecaster receives in a single dataset. As can be seen, under each of the three SPSRs we test, the SSR scores demonstrate a salient linear relationship to the true SPSR scores. We further draw a linear regression curve between the SSR scores and the true scores for each of the three SPSRs of interest (the blue curves in Figure 1). To draw this linear regression curve, we first cluster the score pairs into different groups based on the value of the SPSR scores and compute the center point (the mean score pair) for each group, represented by the orange triangles in Figure 1. Then, we regress on these center points.[7] The three regression curves demonstrate a slope of 0.74, 0.73, and 0.84, respectively, all with an intercept near 0. This result indicates that though the SSR scores are not exactly unbiased in real data, they still follow an affine transformation of the true SPSR scores with decent approximate unbiasedness.

We also notice that under all three SPSRs, the SSR scores tend to underestimate the true scores by around 20%. As the SSR scores follow an affine transformation of the SPSR scores empirically, this underestimation can possibly be mitigated by applying a constant scaling factor (e.g., 1.25 as suggested by our regression) without influencing the incentive properties of the SSR mechanisms.

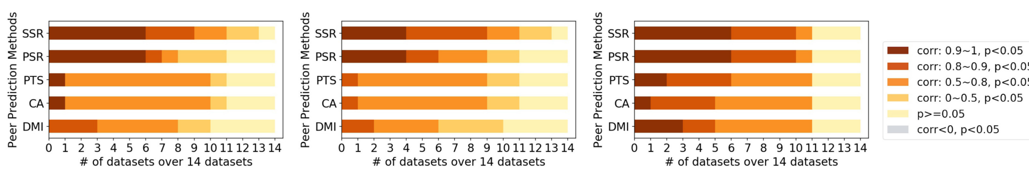
Correlation with SPSR. We compare the correlations between agents' SPSR scores and the scores given by the five peer prediction mechanisms we test. We first measure the correlations on each dataset independently using Pearson's correlation coefficient (corr) and then classify them into different levels based on the value of the coefficient. Finally, we count the number of datasets at

[7]The reason for clustering score pairs before regression is that the SPSR scores of forecasters are not distributed evenly within the range of the SPSR score, with most forecasters' SPSR scores falling in the low range of the SPSR score. Consequently, drawing the regression curve directly on all score pairs will mainly reflect the regression pattern in the low range of the SPSR score instead of the whole range. In fact, for each of the three SPSR tested, the corresponding SSR mechanism obtains a regression slope closer to 1 at the low range of the SPSR score.



(a) Brier ($y = 0.744 \cdot x + 0.014$) (b) Log ($y = 0.733 \cdot x - 0.041$) (c) Rank-sum ($y = 0.839 \cdot x - 0.041$)

Fig. 1. Regression of individuals' true accuracy and SSR score over 14 datasets under 3 different SPSRs.



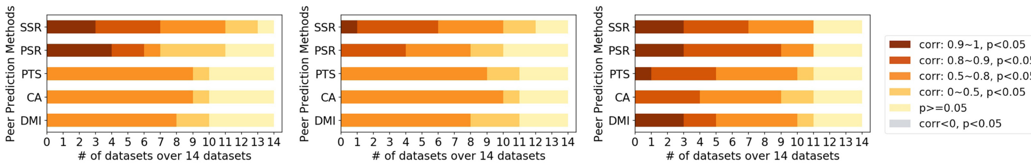
(a) Brier score (b) Log scoring rule (c) Rank-sum scoring rule

Fig. 2. The number of datasets in each level of correlation (measured by Pearson's correlation coefficient) between individuals' peer prediction scores and different SPSRs.

different correlation levels for each peer prediction mechanism and present the results in Figure 2. As can be seen, all five peer prediction mechanisms achieve a strong correlation (corr > 0.5) to the SPSRs on half of the 14 datasets, while the SSR mechanisms demonstrate an even stronger correlation pattern. In particular, the SSR mechanisms achieve a very strong correlation (corr > 0.8) on 9 out of the 14 datasets under all 3 SPSRs, and achieve correlations in more datasets than other mechanisms for each of the following levels: corr > 0.9, corr > 0.8, and corr > 0.5. The advantage of SSRs in the correlation to SPSRs is most salient under the Brier score and is more salient when compared with the PTS, CA, and DMI mechanisms than compared with the PSR mechanisms. We observe similar results using Spearman's rank correlation test (Figure 3), which implies that SSR mechanisms also rank the agents similarly to SPSRs.

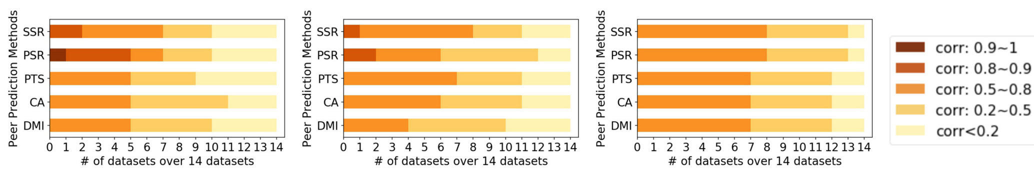
The performance of SSR mechanisms in reflecting the true SPSR scores depends on the accuracy of estimating the error rates of the constructed noisy estimate of ground truth in SSR mechanisms. This estimation accuracy depends on the number of prediction samples that SSR mechanisms have access to. In our previous experiments, each task receives a considerable number of predictions (no less than 20 on average), which may give an edge to the SSR mechanisms. However, a principal with a limited budget can often collect only a small number of predictions for each task. Therefore, we are also interested in comparing the performance of SSR mechanisms with other peer prediction mechanisms when each task receives only a limited number of predictions. To simulate this scenario, for each original dataset, we sample a subset of users to create a new dataset such that each new dataset has an average of 4~5 predictions per task with a minimum of 3 predictions, which is the minimum number of predictions per task required by our SSR mechanisms.[8] Figure 4

[8]To ensure a minimum of 3 predictions per task, we removed a small number of tasks that receive less than 3 predictions by this sampling method. Over the 100 runs of random sampling, around 20 tasks are removed on average from each GJP dataset, and no more than 2 tasks are removed from each of the other datasets in each run. This sampling operation keeps a decent number of predictions for each agent, which allows a stable computation for the scores of SSR, PTS, CA, and DMI mechanisms.



(a) Brier score (b) Log scoring rule (c) Rank sum scoring rule

Fig. 3. The number of datasets in each level of correlation (measured by Spearman's correlation coefficient) between individuals' peer prediction scores and different SPSRs.



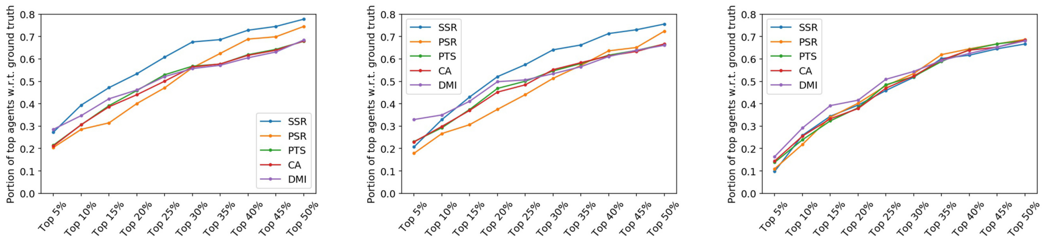
(a) Brier score (b) Log scoring rule (c) Rank-sum scoring rule

Fig. 4. The number of datasets in each level of correlation (measured by Pearson's correlation coefficient) between individuals' peer prediction scores and different SPSRs on sampled datasets (the correlation is averaged over 100 runs of random sampling).

shows the correlation results of each peer prediction mechanism based on the average Pearson's correlation coefficient over 100 runs of random sampling. As can be seen, overall, the correlations between each peer prediction mechanism and the three SPSR in these sampled datasets decrease when compared with the corresponding correlations in the original datasets. SSR mechanisms still maintain a strong correlation (corr > 0.5) over half of the 14 datasets, while the other mechanisms do not. However, the performance difference of SSR and other mechanisms shrinks. The PSR mechanisms outperform SSR mechanisms at two correlation levels, corr > 0.8 and corr > 0.9, under the Brier score and the log scoring rule. In fact, the single-task PSR mechanisms demonstrate smaller correlation decreases, indicating that they are more robust to the number of predictions than the other four multi-task mechanisms.

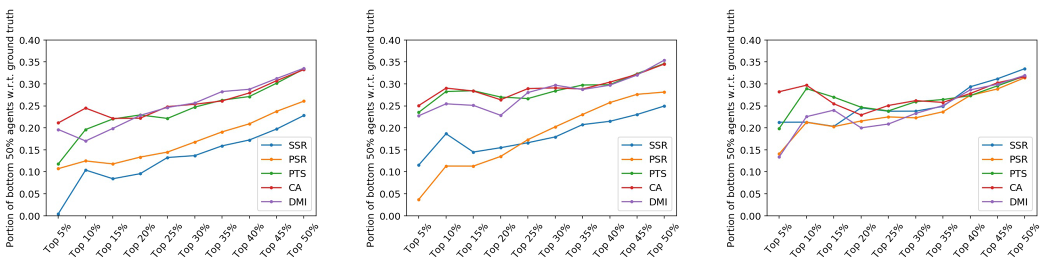
Expert identification. SPSRs are sometimes used to identify top forecasters to assign prizes, for example, in GJP and HFC. Moreover, accurate identification of true top forecasters without access to the ground truth can help improve the aggregation accuracy when a principal wants to aggregate forecasters' predictions into a final prediction for each task [45]. Therefore, we examine to what extent different peer prediction scores can identify top-performing experts in terms of the true SPSRs without access to the ground truth.

In GJP and HFC, forecasters may answer different sets of forecasting questions. It is non-trivial to compare forecasters' performance when they answer different sets of questions whose difficulty levels vary. Here, we use the mean peer prediction scores and mean SPSR scores to measure the forecaster's performance for simplicity, as our main purpose is to compare how close the evaluation results will be when we use true SPSRs and when we use the SSR scores in the same way. Though we are demonstrating the application of SSR in expert identification, it is a by-product of its calibration property and we acknowledge that the identification might be affected by other factors, including how agents selected the forecasting questions. In projects such as GJP, the organizers impute and standardize the scores for different questions and then prize the forecasters.



(a) Brier score (b) Log scoring rule (c) AUC-ROC

Fig. 5. The portion of top $t\%$ forecasters with regard to 3 different metrics (mean squared loss, cross-entropy loss, AUC-ROC loss) in the top $t\%$ forecasters selected by different methods (averaged over 14 datasets).



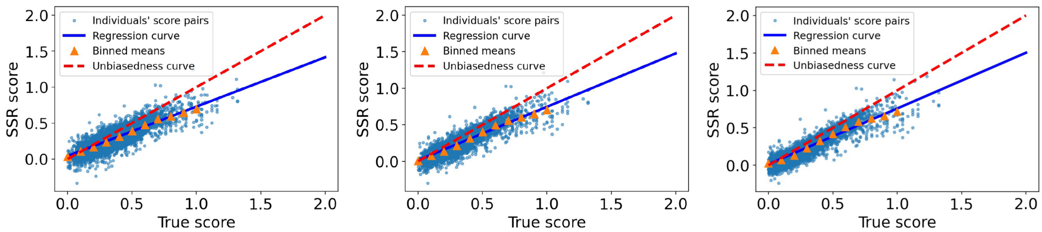
(a) Brier score (b) Log scoring rule (c) AUC-ROC

Fig. 6. The portion of bottom 50% forecasters with regard to 3 different metrics (mean squared loss, cross-entropy loss, AUC-ROC loss) in the top $t\%$ users selected by different methods (averaged over 14 datasets).

Arguably, forecasters have no incentive to choose easy questions *a priori*, which alleviates the evaluation bias induced by heterogeneous difficulty levels of the forecasting questions.

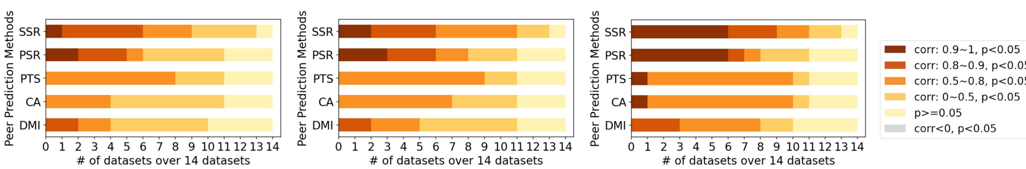
We first rank the forecasters according to one of the three SPSRs (when the rank-sum scoring rule is chosen, we use the AUC-ROC instead to evaluate agents' true accuracy because, as an accuracy metric instead of an incentive device, AUC-ROC is much more popular than the rank-sum scoring rule). We focus on two metrics about expert identification: (i) the percentage of top $t\%$ forecasters identified by the SPSRs in the top $t\%$ forecasters selected by a peer prediction method, and (ii) the percentage of below-average forecasters (the bottom 50% forecasters) under the SPSRs in the top $t\%$ forecasters selected by a peer prediction method. The results are shown in Figures 5 and 6. For illustration, consider the left (Brier) pane of Figure 5. Of the top 10% forecasters according to SSR, 40% are indeed in the top 10% according to the Brier score. We find that for both the Brier score and the log score, there are more true top $t\%$ forecasters in the top $t\%$ forecasters selected by SSR than in the top $t\%$ forecasters selected by other peer prediction mechanisms, when $t\%$ ranges from 5% to 50%. Meanwhile, there are less true below-average forecasters in the top $t\%$ forecasters under SSR and PSR mechanisms than under the other peer prediction mechanisms. For AUC-ROC, while the SSR mechanism maintains a relatively smaller number of true below-average forecasters in its top 10% to 15% forecasters, all 5 peer prediction mechanisms perform similarly, which echoes the correlation results under the rank-sum scoring rule, in which the 5 peer prediction mechanisms all achieve strong correlation in most of the datasets (Figure 2(c)).

Robustness of SSRs in temporal forecasting. The experiments just discussed use the final forecasts of each participant in the two temporal forecasting projects, GJP and HFC. In this experiment, we test whether our evaluation is robust to the choice of the prediction time in temporal



(a) Initial ($y = 0.687x + 0.041$) (b) Average ($y = 0.731x + 0.014$) (c) Final ($y = 0.744 \cdot x + 0.014$)

Fig. 7. Regression and comparison of individuals' mean Brier scores and SSR scores over 14 datasets when the initial, average, and final predictions are used for GJP and HFC datasets.



(a) Initial Prediction (b) Average Prediction (c) Final Prediction

Fig. 8. The number of datasets in each level of correlation (measured by Pearson's correlation coefficient) between individuals' peer prediction scores and the Brier score when the initial, average and final predictions are used for GJP and HFC datasets.

forecasting. In particular, we focus on the Brier score metric and redo the experiments shown in Figures 1(a) and 2(a) while using the initial prediction and the average prediction (from the open date to the close date of the forecasting question) of each participant instead of the final prediction (see Figures 7 and 8).[9] Figure 7 shows that the Pearson's correlation coefficient between SSRs and the Brier score decreases only slightly when the initial and average predictions are used. Figure 8 shows that when we use earlier predictions, the correlations between the five peer prediction scores and the Brier score slightly decrease, while SSRs still maintain a relative advantage in correlation over other mechanisms.

9 DISCUSSION

In this article, we propose the SSR mechanisms such that truthful reporting one's posterior belief is a dominant strategy in the multi-task IEWV setting when each agent uses a consistent reporting strategy across all tasks. Moreover, the reward of a prediction given by an SSR mechanism quantifies the value of information in expectation as if the prediction is assessed by the corresponding SPSR with access to the ground truth. Because of these two properties, our mechanisms are particularly suitable for information elicitation scenarios in which using SPSRs to reward agents is favored but the ground truth is not available in time, such as forecasting long-term geopolitical events and predicting the replicability of social science studies.

There are also some limitations of applying our models and mechanisms. First, our Assumption 2 requires agents' signals on a task to be independent conditioned on the ground truth Y. This implies that our SSR mechanisms apply only to scenarios in which there is such an objective ground truth or in which there is no objective ground truth but the agents' signals are correlated only through a single latent variable. An example of the latter is asking an agent how likely an essay is well

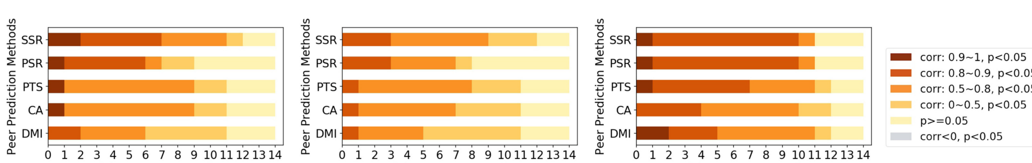
[9]Figures 1(a) and 7(c) are the same. Figures 2(a) and 8(c) are the same.

written or not. Although whether an essay is well written may not have an objective answer, as long as the agents' signals are independent conditioned on a latent variable that captures the real quality of the essay, our mechanisms should incentivize truthful reporting as a dominant strategy when all agents adopt uniform strategies across tasks. In comparison, most existing multi-task peer prediction mechanisms (e.g., [20, 35, 42]) that elicit categorical signals do not require agents' signals to be correlated only through a latent variable. Instead, they allow a broader correlation pattern (e.g., self-predicting [35]) or arbitrary correlations as long as signals are not completely independent (e.g., [20, 42]).

Second, to estimate the error rates of the noisy estimate of the ground truth, our mechanisms require at least three reports for each task. In contrast, several multi-task mechanisms (e.g., [6, 20, 35, 42]) need only one peer agent to achieve their incentive properties. Moreover, the variance of the rewards of SSR mechanisms depends on the number of tasks and reports that the mechanisms have access to. A relatively large number of tasks and reports is needed to obtain a low-variance reward for each agent. As can be seen from our empirical study, although SSR mechanisms still maintain better correlations to the true SPSR scores than the other mechanisms when there are only a few reports per task, SSR mechanisms have a more salient correlation decrease when compared with the case in which each task receives a sufficient number of answers. SSR mechanisms are more sensitive to the size of the dataset. However, our analysis suggests that as long as agents adopt uniform strategies across tasks, it is possible to learn the statistical patterns of agents' reports without influencing the incentive. Therefore, a future direction to mitigate SSR mechanisms' sensitivity to the amount of data is to develop or adopt more sophisticated estimation algorithms that require fewer tasks and reports to achieve stable performance.

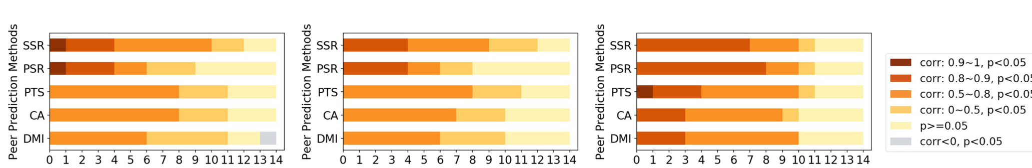
APPENDICES

A MISSING FIGURES



(a) Brier score (b) Log scoring rule (c) Rank-sum scoring rule

Fig. A.1. The number of datasets in each level of correlation (measured by Pearson's correlation coefficient) between individuals' peer prediction scores and different SPSR when each probabilistic prediction is mapped to the most likely binary vote with uniform random tie breaking.



(a) Brier score (b) Log scoring rule (c) Rank sum scoring rule

Fig. A.2. The number of datasets in each level of correlation (measured by Spearman's correlation coefficient) between individuals' peer prediction scores and different SPSRS when each probabilistic prediction is mapped to the most likely binary vote with uniform random tie breaking.

B PROOF OF LEMMA 5.2

Proof. Suppose that z and y are not stochastically relevant; we have that

$$\Pr[y=0|z=0]=\Pr[y=0|z=1], \quad (18)$$
$$\Pr[y=1|z=0]=\Pr[y=1|z=1]. \quad (19)$$

From Equation (18), we know that

$$\frac{\Pr[y=0,z=1]}{\Pr[z=1]}=\frac{\Pr[y=0,z=0]}{\Pr[z=0]} \Leftrightarrow \frac{\Pr[y=0]e_z^-}{\Pr[z=1]}=\frac{\Pr[y=0](1-e_z^-)}{\Pr[z=0]},$$

When $\Pr[y=0]\neq 0$, we hav that $\frac{\Pr[z=1]}{\Pr[z=0]}=\frac{e_z^-}{1-e_z^-}$. Similarly, from Equation (19), we have that $\frac{\Pr[z=1]}{\Pr[z=0]}=\frac{1-e_z^+}{e_z^+}$, when $\Pr[y=1]\neq 0$. Therefore, we obtained that

$$\frac{e_z^-}{1-e_z^-}=\frac{1-e_z^+}{e_z^+},$$

from which we have that $e_z^-+e_z^+=1$. Contradiction. The other direction follows similarly. □

C PROOF OF LEMMA 6.6

Proof. We consider the estimation of the error rates e_z^+, e_z^- of an agent i, and we consider a generic task as tasks are *a priori* similar. Thus, in the proof, we drop the subscript k, which indexes the tasks. There are two layers of estimation error in solving the system of Equations (4), (5), and (6):

- **(1) Estimation error due to heterogeneous agents:** The higher-order equations do not capture the true matching probability with heterogeneous agents. As we draw Z_2 and Z_3 in a task without replacement, with a finite number of agents, Z_2 and Z_3 are dependent on Z_1, and the error rates of Z_2 and Z_3 are not exactly the same as the error rates of Z_1 (z).
- **(2) Estimation errors due to finite estimation samples:** The last sources of errors come from the estimation errors of $\widetilde{\beta_{-i}}, \widetilde{\gamma_{-i}}$, and $\widetilde{\alpha_{-i}}$.

Next, we bound the two errors separately.

(1) Estimation error due to heterogeneous agents: The challenge lies in the fact that the higher-order equations do not capture the true matching probability with heterogeneous agents.

We first consider Equation (5), which is not precise – randomly picking a prediction signal from all agents without replacement leads to different error rates. This will complicate the solution for the system of equations. We show that our estimation, though ignoring the above bias, will not affect our results by much: Let k_1 be the agent whose prediction signal is picked to be Z_1. Conditioned on agent k_1 being picked and on reports $q_1, \ldots, q_N$, we have that $\Pr[Z_1=Z_2=1|q_1,\ldots,q_N,k_1]=q_{k_1}\cdot(\frac{\sum_{j\neq i,k_1} q_j}{N-2})$. Recall that q_{k_1} is a random variable because of the private signal c_{k_1} received by agent k_1 and the randomness in σ_{k_1}, and that $e_z^+=\mathbb{E}_{q_1,\ldots,q_N|y=1}[\bar{q}_{-i}]$. We have that

$$\begin{aligned}
\Pr[Z_1 = Z_2 = 1|y = 1] &= \mathbb{E}_{k_1}[\mathbb{E}_{q_1,\dots,q_N|y=1}[\Pr[Z_1 = Z_2 = 1|k_1, q_1, \dots, q_N]]] \\
&= \mathbb{E}_{k_1}\left[\mathbb{E}_{q_1,\dots,q_N|y=1}\left[q_{k_1} \cdot \left(\frac{\sum_{j\neq i,k_1} q_j}{N-2}\right)\right]\right] \\
&= \mathbb{E}_{k_1}\left[\mathbb{E}_{q_1,\dots,q_N|y=1}[q_{k_1}] \cdot \mathbb{E}_{q_1,\dots,q_N|y=1}\left[\frac{\sum_{j\neq i,k_1} q_j}{N-2}\right]\right] \\
&= \mathbb{E}_{k_1}\left[\mathbb{E}_{q_1,\dots,q_N|y=1}[q_{k_1}] \cdot \mathbb{E}_{q_1,\dots,q_N|y=1}\left[\frac{(N-1)\bar{q}_{-i}}{N-2} - \frac{q_{k_1}}{N-2}\right]\right] \\
&= \mathbb{E}_{k_1}\left[\mathbb{E}_{q_1,\dots,q_N|y=1}[q_{k_1}] \cdot \left(\frac{N-1}{N-2}e_z^+ - \frac{1}{N-2}\mathbb{E}_{q_1,\dots,q_N|y=1}[q_{k_1}]\right)\right] \\
&= \frac{N-1}{N-2}e_z^+\mathbb{E}_{k_1}\left[\mathbb{E}_{q_1,\dots,q_N|y=1}[q_{k_1}]\right] - \frac{1}{N-2}\mathbb{E}_{k_1}\left[\mathbb{E}^2_{q_1,\dots,q_N|y=1}[q_{k_1}]\right] \\
&= \frac{N-1}{N-2}(e_z^+)^2 - \frac{1}{N-2}\omega,
\end{aligned}$$

where $\omega := \mathbb{E}_{k_1}[\mathbb{E}^2_{q_1,\dots,q_N|y=1}[q_{k_1}]]$.

Note that both e_z^+ and ω are no more than 1. Then,

$$\left|\frac{N-1}{N-2}(e_z^+)^2 - \frac{1}{N-2}\omega - (e_z^+)^2\right| \leq \frac{(e_z^+)^2}{N-2} + \frac{1}{N-2}\omega \leq \frac{2}{N-2}$$

This adds $\frac{2}{N-2}$ error bias in the step where we replace $\Pr[z_1 = z_2 = 1|y = 1]$ with $(e_z^+)^2$ in the deduction of Equation (5). In addition, it finally adds $\frac{2}{N-2}$ error bias in estimating β_{-i} (through both $(e_z^-)^2$ and $(1 - e_z^+)^2$) in Equation (5).

Similarly for the matching among three agents (Equation (6)) we have that

$$\left|\Pr[Z_1 = Z_2 = Z_3 = 1|y = 1] - (e_z^+)^3\right| \leq \frac{3}{N-3}.$$

This adds $\frac{3}{N-3}$ error bias in estimating γ_{-i}.

(2) Estimation errors due to finite estimation samples: The last sources of errors come from the estimation errors of $\widetilde{\beta_{-i}}, \widetilde{\gamma_{-i}}$, and $\widetilde{\alpha_{-i}}$. Direct application of the Chernoff bound gives us the following lemma:

Lemma C.1. *When there are M samples for estimating $\widetilde{\beta_{-i}}, \widetilde{\gamma_{-i}}$. and $\widetilde{\alpha_{-i}}$, respectively (total budgeting 3M), we have with probability at least $1-\delta$ that*

$$|\widetilde{\beta_{-i}} - \beta_{-i}| \leq \sqrt{\frac{\ln\frac{6}{\delta}}{2M}},\ |\widetilde{\gamma_{-i}} - \gamma_{-i}| \leq \sqrt{\frac{\ln\frac{6}{\delta}}{2M}},\ |\widetilde{\alpha_{-i}} - \alpha_{-i}| \leq \sqrt{\frac{\ln\frac{6}{\delta}}{2M}}.$$

The error analysis in **1** and **2** jointly imply that with probability at least $1-\delta$

$$|\widetilde{\beta_{-i}} - \beta_{-i}| \leq \sqrt{\frac{\ln\frac{6}{\delta}}{2M}} + \frac{2}{N-2},\ |\widetilde{\gamma_{-i}} - \gamma_{-i}| \leq \sqrt{\frac{\ln\frac{6}{\delta}}{2M}} + \frac{3}{N-3},\ |\widetilde{\alpha_{-i}} - \alpha_{-i}| \leq \sqrt{\frac{\ln\frac{6}{\delta}}{2M}}.$$

Now we are ready to prove Lemma 6.6. First, from Algorithm 1, we can easily derive that

$$|\widetilde{e_z^-} - e_z^-| \leq \frac{|\tilde{a} - a|}{2} + \frac{|\sqrt{\tilde{a}^2 - 4\tilde{b}} - \sqrt{a^2 - 4b}|}{2} \tag{20}$$

$$|\widetilde{e_z^+} - e_z^+| \leq \frac{|\tilde{a} - a|}{2} + \frac{|\sqrt{\tilde{a}^2 - 4\tilde{b}} - \sqrt{a^2 - 4b}|}{2}. \tag{21}$$

For the latter term in Equations (20) and (21), we have that

$$\begin{aligned}\frac{|\sqrt{\tilde{a}^2-4\tilde{b}}-\sqrt{a^2-4b}|}{2} &= \frac{|(\sqrt{\tilde{a}^2-4\tilde{b}}-\sqrt{a^2-4b})\cdot(\sqrt{\tilde{a}^2-4\tilde{b}}+\sqrt{a^2-4b})|}{2(\sqrt{\tilde{a}^2-4\tilde{b}}+\sqrt{a^2-4b})}\\ &\le \frac{|\tilde{a}^2-a^2|}{2\sqrt{a^2-4b}}+\frac{4|\tilde{b}-b|}{2\sqrt{a^2-4b}}\\ &\le \frac{|\tilde{a}-a|^2}{2\sqrt{a^2-4b}}+\frac{a\cdot|\tilde{a}-a|}{\sqrt{a^2-4b}}+\frac{2|\tilde{b}-b|}{\sqrt{a^2-4b}}\end{aligned}$$

The first inequality is due to the fact that we drop the positive $2\sqrt{\tilde{a}^2-4\tilde{b}}$ in the denominator. For the second inequality, note that a is non-negative as, essentially, $a = 1 - e_z^+ + e_z^-$ as shown in the proof for Theorem 6.3.

To summarize, we have that

$$|\widetilde{e_z^-}-e_z^-| \le \left(\frac{1}{2}+\frac{a}{\sqrt{a^2-4b}}\right)|\tilde{a}-a| + \frac{2|\tilde{b}-b|}{\sqrt{a^2-4b}} + \frac{1}{2\sqrt{a^2-4b}}|\tilde{a}-a|^2 \tag{22}$$

$$|\widetilde{e_z^+}-e_z^+| \le \left(\frac{1}{2}+\frac{a}{\sqrt{a^2-4b}}\right)|\tilde{a}-a| + \frac{2|\tilde{b}-b|}{\sqrt{a^2-4b}} + \frac{1}{2\sqrt{a^2-4b}}|\tilde{a}-a|^2. \tag{23}$$

The key tasks here reduce to bounding $|\tilde{a}-a|$ and $|\tilde{b}-b|$. Recall that

$$\begin{aligned}a &:= \frac{\gamma_{-i}-\alpha_{-i}\beta_{-i}}{\beta_{-i}-(\alpha_{-i})^2}\\ b &:= \frac{\alpha_{-i}\gamma_{-i}-(\beta_{-i})^2}{\beta_{-i}-(\alpha_{-i})^2}.\end{aligned}$$

We know the following facts:

$$\begin{aligned}|(\widetilde{\beta_{-i}}-(\widetilde{\alpha_{-i}})^2)-(\beta_{-i}-(\alpha_{-i})^2)| &\le |(\widetilde{\alpha_{-i}})^2-(\alpha_{-i})^2| + |\widetilde{\beta_{-i}}-\beta_{-i}|\\ &\le 2|\widetilde{\alpha_{-i}}-\alpha_{-i}| + |\widetilde{\beta_{-i}}-\beta_{-i}| \le 3\sqrt{\frac{\ln\frac{6}{\delta}}{2M}}+\frac{2}{N-2},\end{aligned}$$

$$\begin{aligned}|(\widetilde{\gamma_{-i}}-\widetilde{\beta_{-i}}\widetilde{\alpha_{-i}})-(\gamma_{-i}-\beta_{-i}\alpha_{-i})| &\le |\widetilde{\gamma_{-i}}-\gamma_{-i}| + |\widetilde{\beta_{-i}}\widetilde{\alpha_{-i}}-\beta_{-i}\alpha_{-i}|\\ &\le |\widetilde{\gamma_{-i}}-\gamma_{-i}| + |\widetilde{\beta_{-i}}-\beta_{-i}| + |\widetilde{\alpha_{-i}}-\alpha_{-i}|\\ &\le 3\sqrt{\frac{\ln\frac{6}{\delta}}{2M}}+\frac{2}{N-2}+\frac{3}{N-3},\end{aligned}$$

$$\begin{aligned}|(\widetilde{\alpha_{-i}}\widetilde{\gamma_{-i}}-(\widetilde{\beta_{-i}})^2)-(\alpha_{-i}\gamma_{-i}-(\beta_{-i})^2)| &\le |\widetilde{\alpha_{-i}}-\alpha_{-i}| + |\widetilde{\gamma_{-i}}-\gamma_{-i}| + 2|\widetilde{\beta_{-i}}-\beta_{-i}|\\ &\le 4\sqrt{\frac{\ln\frac{6}{\delta}}{2M}}+\frac{2}{N-2}+\frac{3}{N-3}.\end{aligned}$$

Next, denoting $\eta = p(1-p)(1-e_z^+ - e_z^-)^2$ (which also means that $\Delta = p(1-p)(x^- - x^+)^2$), we have that

$$\begin{aligned}\beta_{-i} - (\alpha_{-i})^2 &= (1-p)\cdot(x^-)^2 + p\cdot(x^+)^2 - ((1-p)\cdot x^- + p\cdot x^+)^2 \\ &= (1-p)\cdot p\cdot(x^- - x^+)^2 \\ &= \eta.\end{aligned}$$

Let N be sufficiently large such that

$$3\sqrt{\frac{\ln\frac{6}{\delta}}{2M}} + \frac{2}{N-2} < \eta. \tag{24}$$

Then,

$$\widetilde{\beta_{-i}} - (\widetilde{\alpha_{-i}})^2 \geq \frac{p(1-p)}{2}\cdot\frac{\eta}{2}.$$

Therefore,

$$\begin{aligned}|\tilde{a} - a| &= \left|\frac{\widetilde{\gamma_{-i}} - \widetilde{\alpha_{-i}}\widetilde{\beta_{-i}}}{\widetilde{\beta_{-i}} - (\widetilde{\alpha_{-i}^2})} - \frac{\gamma_{-i} - \alpha_{-i}\beta_{-i}}{\beta_{-i} - (\alpha_{-i})^2}\right| \\ &\leq \frac{|(\widetilde{\beta_{-i}} - (\widetilde{\alpha_{-i}})^2) - (\beta_{-i} - (\alpha_{-i})^2)| + |(\widetilde{\gamma_{-i}} - \widetilde{\beta_{-i}}\widetilde{\alpha_{-i}}) - (\gamma_{-i} - \beta_{-i}\alpha_{-i})|}{|\widetilde{\beta_{-i}} - (\widetilde{\alpha_{-i}})^2|\cdot|\beta_{-i} - (\alpha_{-i})^2|} \\ &\leq \frac{2}{\eta^2}\left(6\sqrt{\frac{\ln\frac{6}{\delta}}{2M}} + \frac{4}{N-2} + \frac{3}{N-3}\right).\end{aligned}$$

Note that the first inequality uses Lemma 6.7.

Similarly for b, we have that

$$|\tilde{b} - b| \leq \frac{2}{\eta^2}\left(7\sqrt{\frac{\ln\frac{6}{\delta}}{2M}} + \frac{4}{N-2} + \frac{3}{N-3}\right)$$

Together, we proved that when M and N are sufficiently large such that Equation (24) holds, that is, $3\sqrt{\frac{\ln\frac{6}{\delta}}{2M}} + \frac{2}{N-2} < \frac{\eta}{2}$, we have th

$$|\widetilde{e_z^-} - e_z^-| \leq O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{M}} + \frac{1}{N}\right)$$

$$|\widetilde{e_z^+} - e_z^+| \leq O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{M}} + \frac{1}{N}\right) \qquad \square$$

ACKNOWLEDGMENTS

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, ODNI, IARPA, DARPA, SSC Pacific or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

We also thank the anonymous reviewers of TEAC for constructive feedback that helped us better present the results in the paper. A conference version of the paper has been published in EC'20 with the same title. See https://dl.acm.org/doi/abs/10.1145/3391403.3399488.

REFERENCES

[1] Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave, and Colin Camerer. 2019. Predicting the replicability of social science lab experiments. *PLOS One* 14, 12 (2019).

[2] Dana Angluin and Philip Laird. 1988. Learning from noisy examples. *Machine Learning* 2, 4 (1988), 343–370.

[3] Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. 2016. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science* 63, 3 (2016), 691–706.

[4] Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78, 1 (1950), 1–3.

[5] Tom Bylander. 1994. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the 7th Annual Conference on Computational Learning Theory*. ACM, 340–347.

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

[7] Darrell Duffie and Jun Pan. 1997. An overview of value at risk. *Journal of Derivatives* 4, 3 (1997), 7–49.

[8] Alexander Frankel and Emir Kamenica. 2019. Quantifying information and uncertainty. *American Economic Review* 109, 10 (2019), 3650–80.

[9] Benoît Frénay and Michel Verleysen. 2014. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems* 25, 5 (2014), 845–869.

[10] Jeffrey A. Friedman, Joshua D. Baker, Barbara A. Mellers, Philip E. Tetlock, and Richard Zeckhauser. 2018. The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly* 62, 2 (2018), 410–422.

[11] Rafael Frongillo and Jens Witkowski. 2016. A geometric method to construct minimal peer prediction mechanisms. In *13th AAAI Conference on Artificial Intelligence*.

[12] Tilmann Gneiting and Adrian E. Raftery. 2005. Weather forecasting with ensemble methods. *Science* 310, 5746 (2005), 248–249.

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

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

[15] Suzanne Hoogeveen, Alexandra Sarafoglou, and Eric-Jan Wagenmakers. 2019. Laypeople can predict which social science studies replicate. (2019).

[16] IARPA. 2019. Hybrid Forecasting Competition. Retrieved October 13, 2022 from https://www.iarpa.gov/index.php/research-programs/hfc?id=661.

[17] Victor Richmond Jose, Robert F. Nau, and Robert L. Winkler. 2006. Scoring Rules, Generalized Entropy and Utility Maximization. (2006). Working Paper, Fuqua School of Business, Duke University.

[18] Radu Jurca and Boi Faltings. 2007. Collusion-resistant, incentive-compatible feedback payments. In *Proceedings of the 8th ACM Conference on Electronic Commerce*. ACM, 200–209.

[19] Radu Jurca and Boi Faltings. 2009. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research* 34 (2009), 209–253.

[20] Yuqing Kong. 2020. Dominantly truthful multi-task peer prediction with a constant number of tasks. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2398–2411.

[21] Yuqing Kong, Katrina Ligett, and Grant Schoenebeck. 2016. Putting peer prediction under the micro (economic) scope and making truth-telling focal. In *International Conference on Web and Internet Economics*. Springer, 251–264.

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

[23] Yuqing Kong and Grant Schoenebeck. 2019. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. *ACM Transactions on Economics and Computation* 7, 1 (2019), 2.

[24] Yuqing Kong, Grant Schoenebeck, Biaoshuai Tao, and Fang-Yi Yu. 2020. Information elicitation mechanisms for statistical estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2095–2102.

[25] Yang Liu and Mingyan Liu. 2015. An online learning approach to improving the quality of crowd-sourcing. In *Proceedings of the 2015 ACM SIGMETRICS* (Portland, Oregon). ACM, New York, NY, 217–230.

[26] John McCarthy. 1956. Measures of the value of information. *PNAS: Proceedings of the National Academy of Sciences of the United States of America* 42, 9 (1956), 654–655.

[27] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. 2015. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*. 125–134.

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

[29] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Advances in Neural Information Processing Systems*. 1196–1204.

[30] Matthew Parry et al. 2016. Linear scoring rules for probabilistic binary classification. *Electronic Journal of Statistics* 10, 1 (2016), 1596–1607.

[31] Dražen Prelec. 2004. A Bayesian truth serum for subjective data. *Science* 306, 5695 (2004), 462–466.

[32] Dražen Prelec, H. Sebastian Seung, and John McCoy. 2017. A solution to the single-question crowd wisdom problem. *Nature* 541, 7638 (2017), 532.

[33] Goran Radanovic and Boi Faltings. 2013. A robust Bayesian truth serum for non-binary signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI'13)*.

[34] Goran Radanovic and Boi Faltings. 2014. Incentives for truthful information elicitation of continuous signals. In *28th AAAI Conference on Artificial Intelligence*.

[35] Goran Radanovic, Boi Faltings, and Radu Jurca. 2016. Incentives for effort in crowdsourcing using the peer truth serum. *ACM Transactions on Intelligent Systems and Technology* 7, 4 (2016), 48.

[36] Blake Riley. 2014. Minimum truth serums with optional predictions. In *Proceedings of the 4th Workshop on Social Computing and User Generated Content (SC'14)*.

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

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

[39] Grant Schoenebeck and Fang-Yi Yu. 2020. Two strongly truthful mechanisms for three heterogeneous agents answering one question. In *International Conference on Web and Internet Economics*. Springer, 119–132.

[40] Clayton Scott. 2015. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *AISTATS*.

[41] Clayton Scott, Gilles Blanchard, Gregory Handy, Sara Pozzi, and Marek Flaska. 2013. Classification with asymmetric label noise: Consistency and maximal denoising. In *COLT*. 489–511.

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

[43] Philip E. Tetlock, Barbara A. Mellers, Nick Rohrbaugh, and Eva Chen. 2014. Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science* 23, 4 (2014), 290–295.

[44] Brendan van Rooyen and Robert C. Williamson. 2015. Learning in the presence of corruption. *arXiv preprint:1504.00091* (2015).

[45] Juntao Wang, Yang Liu, and Yiling Chen. 2019. Forecast aggregation via peer prediction. *arXiv preprint arXiv:1910.03779* (2019).

[46] Robert L. Winkler. 1969. Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association* 64, 327 (1969), 1073–1078.

[47] Jens Witkowski, Pavel Atanasov, Lyle H. Ungar, and Andreas Krause. 2017. Proper proxy scoring rules. In *31st AAAI Conference on Artificial Intelligence*.

[48] J. Witkowski and D. C. Parkes. 2012. Peer prediction without a common prior. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC'12)*. ACM, 964–981.

[49] Jens Witkowski and David C. Parkes. 2012. A robust Bayesian truth serum for small populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*.

[50] Jens Witkowski and David C. Parkes. 2013. Learning the prior in minimal peer prediction. In *Proceedings of the 3rd Workshop on Social Computing and User Generated Content at the ACM Conference on Electronic Commerce*, Vol. 14.

Received 15 August 2020; revised 10 September 2022; accepted 30 September 2022