Surrogate Scoring Rules

YANG LIU*, UC Santa Cruz, USA
JUNTAO WANG*, Harvard University, USA
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 paper, we extend such scoring rules to settings where a principal elicits private probabilistic beliefs but only has access to agents' reports. We name our solution *Surrogate Scoring Rules* (SSR). SSR build on a bias correction step and an error rate estimation procedure for a reference answer defined using agents' reports. We show that, with a single bit of 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 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. 2020. Surrogate Scoring Rules. In *Proceedings of the 21st ACM Conference on Economics and Computation (EC '20), July 13–17, 2020, Virtual Event, Hungary*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3391403.3399488

1 INTRODUCTION

Strictly proper scoring rules (SPSR) [3, 10, 13, 28, 33] have been developed to elicit private information (e.g. probability assessment about whether the S&P 500 index will go up next week) and evaluate the reported information for settings where the principal will have access to the ground truth (e.g. whether S&P 500 index actually went up) at some point. The score of an agent measures the quality of her prediction. Moreover, facing a strictly proper scoring rule, the agent strictly maximizes her expected score by truthfully revealing her prediction. In this paper, we focus on extending the literature of SPSR to the information elicitation *without* verification (IEWV) settings where the principal does not have access to the ground truth and still wants to elicit private probabilistic beliefs. We ask the following question:

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

*Both authors contributed equally to this research.

EC '20, July 13–17, 2020, Virtual Event, Hungary

ACM ISBN 978-1-4503-7975-5/20/07…$15.00
https://doi.org/10.1145/3391403.3399488

We provide a positive answer to this question for multi-task information elicitation. We develop a family of scoring mechanisms that under certain assumptions can estimate a biased version of the ground truth and score predictions against it by removing the bias. As a consequence, we achieve a certain form of dominant truthfulness in eliciting private probabilistic information, a favorable property to have for IEWV [6, 9, 11, 14, 15, 17]. To the best of our knowledge, this is the first work to provide a meta solution framework that enables applications of a SPSR to the IEWV setting for eliciting probabilistic beliefs. We name our solution as *Surrogate Scoring Rules*.

As a building block, we first introduce SSR for a stylized setting where the principal has a noisy ground truth (and its error rates) to evaluate the quality of elicited information. We show that SSR preserve the same information quantification and truthful elicitation properties just as SPSR, despite the lack of access to the exact ground truth. These surrogate scoring rules are inspired by the use of surrogate loss functions in machine learning [1, 4, 22, 29, 30]. They remove bias from the noisy ground truth such that in expectation a report is as if evaluated against the ground truth.

Built upon the above bias correction step, when the principal only has access to agents' reports and one bit of information about the marginal distribution of the ground truth over the entire task set, we develop a multi-agent, multi-task mechanism, *SSR mechanism* , to again achieve information quantification and truthful elicitation under dominant strategy, when agents adopt the same (arbitrary) strategy for all the tasks they are assigned, and when the principal has sufficiently many tasks and agents. The method relies on an estimation procedure to accurately estimate the average bias in the peer agents' reports. With the estimation, a random peer agent's report serves as a noisy ground truth and SSR can then be applied smoothly to achieve the two desired properties.

We evaluate the empirical performance of SSR with 14 real-world human forecast datasets. The results show that SSR effectively recover, from only agents' reports, the true scores of agents given by SPSR with ground truth.

We summarize our contributions as follows:

- We extend Strictly Proper Scoring Rules (SPSR) to a family of scoring mechanisms, *Surrogate Scoring Rules* (SSR), that operate in the information elicitation without verification (IEWV) setting. SSR only require access to peer reports and one-bit information on the prior, and are able to truthfully elicit probabilistic beliefs.
- SSR can build upon any existing SPSR and quantify the accuracy or value of the reported information as the SPSR do. Therefore, our work complements the proper scoring rule literature, and this extension largely expands the application of SPSR in challenging elicitation setting where the ground truth is unavailable.
- For the IEWV setting, a SSR alike mechanism (*SSR mechanism*) induces dominant truthfulness in reporting. To the best of our knowledge, it is the first dominantly truthful mechanism that elicit probabilistic predictions.[1] Therefore, we also contribute to the peer prediction literature via providing a mechanism that elicits truthful probabilistic report in *dominant strategy* and rewards agents according to *prediction accuracy w.r.t. SPSR* instead of correlation.
- We evaluate the empirical performance of SSR mechanism on 14 real-world human prediction dataset. The results show that SSR are able to better assess the true accuracy of agents than other existing peer prediction methods.

Organization. The rest of the paper is organized as follows. We survey the most relevant results in the rest of this section. Section 2 lays out the preliminaries. Section 3 provides our model of IEWV. In Section 4, we study the information elicitation problem in the stylized setting, where there is a noisy version of the ground truth with known bias. We introduce surrogate scoring rules

[1]The mechanism proposed in [16] elicits probabilistic predictions but it is not dominantly truthful. The (variants of) mechanisms proposed in [5, 14, 17, 31] are dominantly truthful but they elicit categorical information.

as a powerful solution in this section. In Section 5, we propose the dominantly truthful mechanism, SSR mechanism, to address the general IEWV problem. We present our experimental study about our mechanisms in Section 6. We conclude the paper with Section 7. Missing details and proofs can be found in the Appendix of the full version of this paper [18].

1.1 Related work

The most relevant literature to our paper is *strictly proper scoring rules* and *peer prediction*. SPSR are designed to elicit subjective beliefs of random variables when the principal can evaluate agents' prediction after the random variables realize. The pioneer work [3] proposes the famous Brier score to quantify the quality of forecasts. Works for variants and full characterization results of SPSR include [10, 13, 28, 33].

Peer prediction is the most popular solution to IEWV. Its core idea is to score each agent based on a reference report elicited from the rest of the agents, and to leverage on the stochastic correlation between different agents' information. Earlier peer prediction mechanisms incentivize truthfully reporting at a Bayesian Nash Equilibrium (BNE) [21, 24, 26, 35, 36]. Recent works [5, 15, 31] have made truthful equilibrium focal in the sense that it leads to the highest expected payoff to agents among all equilibria. But there is at least one other equilibrium that gives the same expected payoff to agents. Several more recent works established dominant truthfulness [6, 9, 11, 14, 15, 17]. In particular, [15, 17, 27] achieve truthful reporting in dominant strategy with infinite number of tasks, with the follow-up work [14] achieving this goal with finite tasks.

Most of the peer prediction works focus on eliciting categorical signals instead of probabilistic beliefs. [16] provides a mechanism to elicit probabilistic predictions, but truthfully reporting is an equilibrium strategy instead of a dominant strategy. When the principal does not have the access to the ground truth but an unbiased estimator, [34] develops a family of proper scoring rules that quantifies the value of probabilistic predictions up to an affine transformation [7]. In comparison, our mechanism does not require to know the ground truth or an unbiased estimate, while it elicits truthful probabilistic predictions in dominant strategy, and qualifies the value of information in the predictions as the SPSR does. We emphasize again that our solution SSR provide a meta framework that maps each existing SPSR to a scoring method to elicit continuous probabilistic predictions.

As mentioned, our work borrows ideas from the machine learning literature on learning with noisy data (e.g., [8, 22, 29, 32]). At a high level, our goal in this paper 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.

2 PRELIMINARIES

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

SPSR are designed for eliciting subjective probability distributions of random variables when the principal can reward agents after the realization of the random variables. SPSR apply to eliciting predictions for any random variables, but we introduce them for binary random variables in this section because the rest of our paper focuses on the binary case. Let $y \in \{0, 1\}$ represent a binary event. An agent has subjective belief p for the likelihood of $y = 1$. When the agent reports a prediction q for outcome $y = 1$, the principal rewards the agent using a scoring function $S(q, y)$

that depends on both the agent's report and the realized outcome. Strict properness of $S(\cdot,\cdot)$ is defined as follows.

Definition 2.1. A function $S : [0,1] \times \{0,1\} \to \mathbb{R}$ that maps the reported belief q and the ground truth y into a score is a *strictly proper scoring rule* if it satisfies $\mathbb{E}[S(p,y)] > \mathbb{E}[S(q,y)]$, for all $p, q \in [0,1]$ and $p \neq q$. The expectation is taken with respect to $y \sim \text{Bernoulli}(p)$.

There is a rich family of strictly proper scoring rules, including Brier ($S(q,y) = 1-(q-y)^2$), logarithmic ($S(q,y) = \log(q)$ if $y = 1$ and $S(q,y) = \log(1-q)$ if $y = 0$) and spherical scoring rules [10].

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

SPSR quantify value of information. Another nice property of SPSR is that they quantify the value/accuracy of reported predictions. To give a rigorous argument, we use an indicator vector $\mathbf{y}$ of length 2 to represent outcome 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 [10, 19, 28], any strictly proper scoring rule can be characterized using a corresponding strictly convex function G as follows: $S(q,y) = G(\mathbf{y}) - D_G(\mathbf{y},\mathbf{q})$, where D_G is the Bregman divergence function of G. Now consider the unknown true distribution of y, denoted $\mathbf{p}^* = (1-p^*, p^*)$. The expected score (with respect to $\mathbf{p}^*$) of an agent with prediction q is

$$\mathbb{E}_{y\sim\mathbf{p}^*}[S(q,y)] = \mathbb{E}_{y\sim\mathbf{p}^*}[G(\mathbf{y})] - \mathbb{E}_{y\sim\mathbf{p}^*}[D_G(\mathbf{y},\mathbf{q})].$$

This means that the maximum score an agent can receives in expectation is $\mathbb{E}_{y\sim\mathbf{p}^*}[G(\mathbf{y})]$ and this happens when the agent reports $\mathbf{q} = \mathbf{p}^*$. Moreover, a prediction $\mathbf{q}$ with smaller divergence $\mathbb{E}_{y\sim\mathbf{p}^*}[D_G(\mathbf{y},\mathbf{q})]$ receives higher score in expectation. Intuitively, $\mathbb{E}_{y\sim\mathbf{p}^*}[D_G(\mathbf{y},\mathbf{q})]$ 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 the accuracy of a prediction q based on the corresponding divergence function. When S is taken as the Brier scoring rule, the corresponding Bregman divergence is the quadratic function. Then $\mathbb{E}_{y\sim\mathbf{p}^*}[D_G(\mathbf{y},\mathbf{q})] = ||\mathbf{p}^* - \mathbf{q}||^2$, implying that a prediction 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 relative entropy. Then, $\mathbb{E}_{y\sim\mathbf{p}^*}[D_G(\mathbf{y},\mathbf{q})] = D_{KL}(\mathbf{p}^*||\mathbf{q}) + H(\mathbf{p}^*)$ where H is the entropy function. A prediction with smaller KL-divergence from $\mathbf{p}^*$ receives a higher score in expectation. This property of SPSR allows the principal to take an expert's average score over a set of prediction tasks as a proxy of his average accuracy and rank experts accordingly.

3 OUR MODEL

The goal of this work is to develop scoring mechanisms that quantify the value of elicited information and are incentive compatible, similar to SPSR, but for settings without verification, i.e. when the principal does not have access to the realization of the predicted binary events. We model the information elicitation without verification problem for a multi-task setting. The details of our model and our design goals are described below.

3.1 Model of Information Structure

A principal has a set of $[M] = \{1, ..., M\}$ binary random variables (tasks) $y_k \in \{0,1\}$ for all $k \in [M]$, which she wants to obtain predictions for. Part of our results can be generalized to non-binary tasks, which can be found in Section B of the Appendix [18]. There is a set $[N] = \{1, ..., N\}$ of agents. Neither the principal nor the agents have access to the ground truth y_k, but agents each

observe a private signal $o_{i,k}$, which relates to y_k, for task k, where $o_{i,k}$ comes from a finite domain $[O_i] = \{0, 1, ..., O_i\}$. We allow that the domains of signals differ across agents. We make a few assumptions on the information structure of this setting.

Assumption 1. *Tasks are independent and similar a priori, that is, the joint distribution of* $(o_{1,k}, ..., o_{N,k}, y_k)$ *is i.i.d. for all task* $k \in [M]$.

This assumption is natural when the set of tasks are of similar nature, for example, tasks asking about the reproducibility of studies published in a particular journal within a certain time period. While researchers may a priori hold some beliefs about the journal-wide replication rate, they receive private signals about each study which allows them to give more informed predictions for individual studies. We note that most studies in the field of IEWV make a similar assumption.[2]

Agents share a common prior $p := \Pr[y_k = 1]$ for each task k. We denote the distribution of a signal $o_{i,k}$ conditioned on y_k by $\mathcal{D}_i^+$ (conditioned on $y_k = 1$) and $\mathcal{D}_i^-$ (conditioned on $y_k = 0$). According to Assumption 1, this conditional distribution $(\mathcal{D}_i^+, \mathcal{D}_i^-)$ is shared across different tasks for agent i. We assume that $\mathcal{D}_i^+ \neq \mathcal{D}_i^-$, otherwise, $o_{i,k}$ is independent with y_k. Each agent knows her own $\mathcal{D}_i^+$ and $\mathcal{D}_i^-$. For each task, we further assume that agents' signals are independent conditioned on the ground truth.

Assumption 2. *For each task, the agents' signals are mutually independent conditional on the ground truth. That is,* $\forall k \in [M], \Pr\left[o_{1,k}, ..., o_{N,k} | y_k\right] = \prod_{i \in [N]} \Pr[o_{i,k} | y_k]$.

This assumption is to exclude scenarios where agents have some form of "side information" to coordinate reports. With "side information", it is impossible to have any mechanism that can truthfully elicit agents' predictions without access to the ground truth. This issue has been noted in IEWV for objective questions by Kong et al. [14, 16] and the same assumption has been adopted.

Each agent forms her own belief about y_k based on her received signal $o_{i,k}$. We use $p_{i,k} := \Pr[y = 1 | o_{i,k}]$ to represent agent i's posterior belief on task k. The principal, who knows neither the prior p nor the conditional signal distributions $\mathcal{D}_i^+$ and $\mathcal{D}_i^-$, hopes to elicit predictions $p_{i,k}$ from some agents. We make a technical assumption about the prior and the knowledge of the principal.

Assumption 3. *The common prior* $p \neq 0.5$ *and the principal knows* $\mathbb{1}(p > 0.5)$.

We assume that the principal knows one bit of information about the prior of tasks. This bit of information can help the principal distinguish between a set of truthful predictions vs. a set of inverted predictions (i.e. everyone reporting $1 - p_{i,k}$ instead of $p_{i,k}$), which otherwise is impossible. In practice, this bit of information is usually easy to get. For example, the principal may not know the replication rate of a journal but knows whether on average more than half of the studies are successfully replicated. The assumption $p \neq 0.5$ is a technical condition we will need later to distinguish the true scenario from the inverted one.

$p_{i,k}$ encodes the randomness of $o_{i,k}$. And, $p_{i,k}$ is a discrete random variable with values taken in [0,1]. Assumptions 1 and 2 jointly imply that the agents' posterior beliefs $p_{i,k}$ are homogeneous across tasks and conditionally independent across agents.

Proposition 3.1. *Under Assumptions 1 and 2, agents' beliefs* $p_{i,k}$ *are*

[2] In [5, 14, 17, 27, 31], where they consider information elicitation for subjective questions (i.e., questions with no ground truth concept, e.g., how do you rank the movie), the authors all assumed that the joint distribution of agents' signals is the same for each task and signals are independent across tasks. In [14, 16], where they consider information elicitation for objective questions (i.e., questions with ground truth), the authors all assumed that the joint distribution of agents' signals together with the ground truth is the same for each task, and all signals and the ground truth are independent across tasks.

- *Conditionally homogeneous and independent across tasks: For each agent $i \in [N]$, conditioned on y_k, her posterior beliefs $p_{i,k}$ are i.i.d. for all tasks $k \in [M]$. That is, $\forall k, k' \in [M]$ and $k \neq k'$, $\forall u \in [0,1], \forall v \in \{0,1\}$, $\underline{\Pr[p_{i,k} = u | y_k = v] = \Pr[p_{i,k'} = u | y_{k'} = v]}$; and $\forall M' \subseteq [M]$, $\underline{\Pr[\{p_{i,k}\}_{k \in M'} | \{y_k\}_{k \in M'}] = \prod_{k \in M'} \Pr[p_{i,k} | y_k]}$.*
- *Conditionally independent across agents: $\forall k \in [M], \Pr[p_{1,k}, ..., p_{N,k} | y_k] = \prod_{i \in [N]} \Pr[p_{i,k} | y_k]$.*

The "conditionally homogeneous" condition simply states that agent's "expertise levels" are similar across tasks with same outcomes. In fact, our results hold for models with more general information structures as long as Proposition 3.1 and Assumption 3 are satisfied.[3]

3.2 Mechanism design goals

The principal is interested in designing a scoring mechanism to facilitate the elicitation of predictions for y_k. For each task k, the principal can ask some subset $[N_k] \subseteq [N]$ agents to give a prediction $q_{i,k}, \forall i \in [N_k]$. $q_{i,k}$ can be different from $p_{i,k}$. The principal then pays each agent scores based on the predictions she collects from all tasks. We denote $[M_i] \subseteq M$ the set of tasks agent i answers.

Given a mechanism, an agent may report her belief via some strategy and influence the final predictions elicited. We consider that agents adopt strategies for each task independently, but each strategy could be a mixed strategy.

Definition 3.2. 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 $\sigma : [0,1] \to \Delta_{[0,1]}$ that maps her posterior belief $p_{i,k}$ into a distribution $\sigma(p_{i,k})$ over [0,1] such that the agent draws a report q_i from $\sigma(p_{i,k})$.

We define a strategy as a mapping from the space of posterior beliefs, rather than from the space of private signals. This is without loss of generality because if two realizations of $o_{i,k}$ give the same posterior, we can merge the two realizations into one combined realization in our model. We also assume that each agent adopts the same strategy across tasks.

Assumption 4. *(Consistent Strategy) For any agent $i \in [N]$, she adopts the same strategy $\sigma_i(\cdot)$ over all tasks $k \in [M_i]$.*

This assumption is reasonable as we assume that tasks are a priori similar to each agent. We denote the strategy adopted by agent i on all tasks by $\sigma_i(\cdot)$ and denote the strategy profile of all agents except agent i by σ_{-i}. We also sometimes abuse our notations and use σ_i and σ_{-i} to represent the predictions resulted from these strategies.

The principal would like to design a mechanism $\mathcal{M}$ that, when only having access to the reported predictions of the agents, can score agents for each of their reported predictions. The score that agent i receives for predicting $q_{i,k}$ for task k, when other agents use strategies σ_{-i} on all assigned tasks, is denoted as $R_i(q_{i,k}; \sigma_{-i})$. $R_i(q_{i,k}; \sigma_{-i})$ depends on agent i's prediction on task k and can depend on other agents' predictions on all other tasks. We restrict our attention to anonymous mechanisms and hence drop the subscript i in the score function: we have $R(q_{i,k}; \sigma_{-i})$ as the score of prediction $q_{i,k}$. $\mathbb{E}[R(q_{i,k}; \sigma_{-i})]$ is the expected score that agent i receives for reporting $q_{i,k}$ when other agents use strategies σ_{-i}. The expectation is taken over the randomness in the ground truth, other agents' signals, and other agents' strategies.

In this IEWV setting, the principal hopes to design $\mathcal{M}$ with similar properties as what SPSR have for the information elicitation with verification settings: quantification of the value of information and incentive compatibility.

[3]Here we allow the priors of different tasks to be different and the p in Assumption 3 refers to the mean prior of all tasks.

Quantify value of information. The score of each prediction should reflect the true accuracy of the prediction, similar to what SPSR achieve. That is, for all i, k and $q_{i,k}$ and for any true distribution of ground truth y_k, $\mathbb{E}[R(q_{i,k}; \sigma_{-i})] = f\left(E_{y_k}[S(q_{i,k}, y_k)]\right)$ holds for a SPSR $S(\cdot, \cdot)$ and a strictly increasing function f.

This design goal aspires that the score an agent receives for a prediction in IEWV recovers what the agent would receive with a SPSR (with access to the ground truth) in expectation.

Dominant truthfulness. A mechanism is dominantly truthful if each agent reporting truthfully on each assigned task leads to higher expected payoff than other strategies, regardless of other agents' reporting strategies.

Definition 3.3. For an agent i, a strategy σ_i is a (weakly) dominant strategy if $\forall k \in [M_i]$ and $o_{i,k}$, $\forall i \in [N]$, $\forall \{\mathcal{D}_j^+, \mathcal{D}_j^-\}_{j\in[N]}$, $\forall \sigma_i', \forall \sigma_{-i} : \mathbb{E}[R(\sigma_i; \sigma_{-i})|o_{i,k}] \geq \mathbb{E}[R(\sigma_i'; \sigma_{-i})|o_{i,k}]$, and σ_i is a strictly dominant strategy if the equality holds only when $\sigma_i' = \sigma_i$.

A dominant truthful mechanism in IEVW is a mechanism where truthful reporting is each agent's weakly dominant strategy and a strictly dominant strategy if her peers' reports are informative [4]. Let σ_i^* be the truthful reporting strategy for agent i, i.e., σ_i^* is the function that maps a belief p_i to a distribution where all probability mass is put on p_i. Let $\bar{q}_{-i,k} := \frac{1}{N-1}\sum_{j\neq i} q_{j,k}$ be the mean of agents' reported predictions other than 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 $o_{j,k}$ received by agent j for $j \neq i$. 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 dominantly truthful mechanisms as follows.

Definition 3.4. (Dominant truthfulness). A mechanism $\mathcal{M}$ is *dominantly truthful* if $\forall i \in [N]$, $\forall k \in [M_i]$ and $o_{i,k}$, $\forall \{\mathcal{D}_j^+, \mathcal{D}_j^-\}_{j\in[N]}$, $\forall \sigma_i \neq \sigma_i^*, \forall \sigma_{-i} : \mathbb{E}[R(\sigma_i^*; \sigma_{-i})|o_{i,k}] \geq \mathbb{E}[R(\sigma_i; \sigma_{-i})|o_{i,k}]$, and the inequality holds strictly for any strategy profile σ_{-i} under which $\bar{q}_{-i,k}$ is informative about y_k.

In Definition 3.4, we characterize the condition that peers' reports are informative by that the expectation of the mean of peers' reports differs for different realizations of the ground truth.

4 ELICITATION WITH NOISY GROUND TRUTH

Before we develop mechanisms with desirable properties for our general model, we first achieve these desirable properties, in this section, under a very stylized setting: *elicitation with noisy ground truth*. In this setting, we introduce surrogate scoring rules as an effective solution. These scoring rules will be the building blocks of our mechanisms for the general model.

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 the posterior $p_i = \Pr[y = 1|o_i]$. The principal, although cannot observe y, has access to a noisy ground truth z that has two *error rates*, e_z^+ and e_z^-, defined as follows: $e_z^+ := \Pr[z = 0|y = 1]$, $e_z^- := \Pr[z = 1|y = 0]$. They are the probabilities that z mismatches y under the two realizations of y. The principal knows the realization z and e_z^+, e_z^-. The principal cannot expect to do much if z is independent of y. Hence, we assume that z and y are stochastically relevant, an assumption commonly adopted in the information elicitation literature [21].

[4]Usually, in a dominant truthful mechanism, truthful reporting is the strict dominant strategy. In IEWV, however, if all the peer agents report predictions independently w.r.t. the ground truth, 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 truthfully reporting strictly higher only when the peer reports are informative about the ground truth. For example, in [14, 17], the dominant truthful mechanism is defined to be a mechanism that pays truthful reporting strictly higher when for each agent, there exists at least one peer agent reporting truthfully. We will see later that in our definition, we do not require that at least one peer agent reports truthfully. We allow all peer agents to be non-truthful but the mean of their peers reports should be dependent with the ground truth.

Definition 4.1. Random variable z is stochastically relevant for random variable y if the distribution of y conditioned on z is different for different realizations of z.

The following lemma shows that the stochastic relevance requirement directly translates to a constraint on the error rates, that is, $e_z^+ + e_z^- \neq 1$.

Lemma 4.2. *z is stochastically relevant to 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 using this noisy ground truth z and the knowledge of error rates e_z^+, e_z^-. We define the design space of the scoring rule with noisy ground truth as follows.

Definition 4.3. Given a noisy ground truth z with error rates $(e_z^+, e_z^-) \in [0,1]^2$, *a scoring rule with noisy 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 ground truth $z \in \{0,1\}$ to a score. The function R can depend on error rates (e_z^+, e_z^-).

Adopting the terminology from the scoring rule literature, we refer to strict properness as the property that a scoring rule with noisy ground truth gives a strictly higher expected score to a truthful report than a non-truthful report.

Definition 4.4. A scoring rule $R(q_i, z)$ with noisy ground truth z is *strictly proper* if it holds for all realizations of o_i and $p_i = \Pr[y = 1|o_i]$, that $\forall q_i \in [0,1](q_i \neq p_i), \mathbb{E}_{z|o_i}[R(p_i, z)] > \mathbb{E}_{z|o_i}[R(q_i, z)]$.

4.1 Surrogate scoring rules (SSR)

In this section, we present our solution, the surrogate scoring rules, for this stylized setting. SSR is a family of scoring rules with noisy ground truth and is strictly proper under mild conditions.

Definition 4.5 (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 for that $\forall p_i, q_i, e_z^+, e_z^- \in [0,1]$ and $e_z^+ + e_z^- \neq 1$, $\underline{\mathbb{E}_z[R(q_i, z)] = f\left(\mathbb{E}_y[S(q_i, y)]\right)}$, where y is the ground truth drawn from Bernoulli(p_i) and z is the noisy ground truth generated by y with error rates e_z^+, e_z^-.

The above definition seeks a surrogate scoring rule $R(\cdot)$ that helps us remove the bias in z and return us a strictly proper score in expectation. The idea is borrowed from the machine learning literature on learning with noisy data [4, 20, 22, 29, 32]. SSR can be viewed as a particular class of proxy scoring rules [34]. But the approach of [34] to achieve properness is to plug in an *unbiased* proxy ground truth to a strictly proper scoring rule. SSR on the other hand directly work with biased proxy and the scoring function is designed to de-bias the noise. Easily we have the following strict properness result for SSR:

Theorem 4.6. *Given an agent's fixed prior p and private signal o_i, SSR $R(q_i, z)$ with noisy ground truth 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 are stochastically relevant to y.*

We give an implementation of SSR, which we name as SSR_α :

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

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

where S can be any strictly proper scoring rule. We note that the knowledge of the error rates e_z^+, e_z^- is crucial for defining the above SSR. This SSR function is inspired by Natarajan et al.[22]. It has the following property:

Mechanism 1 SSR mechanism (Sketch)

1: For each task k, we uniformly randomly pick at least 3 agents, assign task k to them and collect their reported predictions.
2: For each agent i and each task k she answers, we construct a reference report $z_{i,k}$ using peer agents' reports; estimate the error rates $e^+_{z_{i,k}}$ and $e^-_{z_{i,k}}$ for $z_{i,k}$.
3: Pay each agent i for $q_{i,k}$ 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.

Lemma 4.7 (Lemma 1, [22]). *For SSR_α : $\forall q_i, e^+_z, e^-_z \in [0,1]$ and $e^+_z + e^-_z \neq 1, \forall y \in \{0,1\}$: $\mathbb{E}_{z|y}[R(q_i, z)] = S(q_i, y)$.*

Intuitively speaking, the linear transform in SSR_α will ensure that in expectation, the prediction q_i is scored as if it was scored against y using a SPSR. This can be proved fairly straightforwardly via spelling out the expectation. Interested readers are also referred to [22]. We would like to note that other surrogate loss functions designed for learning with noisy labels can also be leveraged to design SSR.

Theorem 4.8. *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 ground truth generated by y with error rate e^+_z, e^-_z.*

With Theorem 4.8 we know that SSR_α quantifies the quality of information just as the strictly proper scoring rule S does. Further, SSR_α has the following variance:

Theorem 4.9. *Let $p_z := \Pr[z = 1]$. SSR_α suffers the following variance:*

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

5 ELICITATION WITHOUT VERIFICATION

The results in the previous section are built upon the fact that there exists a noisy copy of the ground truth and we know its error rates. In this section, we apply the idea of SSR to information elicitation without verification. A reasonable way to do so is to take agents' reports as the source for this noisy reference of the ground truth. Yet the principal cannot assume the knowledge of the noise in agents' reports. We find a way to construct a noisy ground truth from agents' report with estimable error rates. We refer this noisy ground truth as the *reference report*. Applying SSR with this reference report, we can finally get a dominantly truthful mechanism that elicits the information and that the payment of the mechanism also quantifies the value of information of agents' reports as what the SPSR do. We call this mechanism *SSR mechanism* . We present the sketch of our mechanism in Mechanism 1.

The challenge of designing such a mechanism is to construct such a 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 such a reference report and how to estimate the error rates.

5.1 Reference report and its property

Let $s_{j,k}$ be a binary signal independently drawn from Bernoulli$(q_{j,k})$. We term $s_{j,k}$ 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 the peer agent set $[N]\backslash\{i\}$, collect her prediction $q_{j,k}$, and draw the prediction signal $s_{j,k} \sim$ Bernoulli$(q_{j,k})$. We use this $s_{j,k}$ as the reference report $z_{i,k}$.*

Clearly, conditioned on the reports $q_{j,k}, j \in [N]$, the distribution of $z_{i,k}$ is Bernoulli$\big(\bar{q}_{-i,k}\big)$ as we uniformly randomly pick a prediction signal. Note that in our model, $q_{i,k} \sim \sigma_i(p_{i,k}), i \in [N], k \in$

$[M]$. Due to Proposition 3.1 and Assumption 4, $\bar{q}_{-i,k}$ is i.i.d. across tasks $k \in [M]$. Thus, $z_{i,k}, k \in [M]$ have the following two properties.

LEMMA 5.1. *$\forall i \in [N], k \in [M]$, $z_{i,k}$ 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 ground truth by Theorem 4.6 and thus, SSR with $z_{i,k}$ is strictly proper for eliciting the posterior belief $p_{i,k}$.

LEMMA 5.2. *For any strategy profile agents play, reference reports of an agent $i \in [N]$ are i.i.d. and have the same error rates w.r.t. the ground truth, i.e., $\forall \sigma_1, ..., \sigma_N, \forall i \in [N], \exists e_i^+, e_i^- \in [0,1], \forall k \in [M]$: $\Pr[z_{i,k} = 0 | y_k = 1] = e_i^+, \Pr[z_{i,k} = 1 | y_k = 0] = e_i^-$.*

This lemma shows that the error rates of the reference reports for agent i are the same across all tasks. This property makes it possible to estimate the error rates using multi-task data. In the following sections, we introduce the estimation of the error rates and complete our mechanism.

5.2 Asymptotic setting

To better deliver our idea for error rates estimation, we start with an asymptotic setting with infinite amounts of tasks and agents, i.e., $M, N \to \infty$. We will later provide finite sample justification for our mechanism.

We focus on estimating the error rates of the reference reports for agent i. Based on Lemma 5.2, we can use z to denote the reference report for agent i on a generic task, and we only need to estimate the error rates e_z^+, e_z^- of z. Our estimation algorithm relies on establishing three equations. We show that the three equations, with knowing their true parameters (which is true in the asymptotic setting), together will uniquely define e_z^+, e_z^-. Then, in next section, we argue that in the finite sample setting, with imperfect estimate of parameters from agents' reports, the solution from the perturbed set of equations will approximate the true values of e_z^+, e_z^-, with guaranteed accuracy.

To construct the three equations, we make the following preparation. Let $\underline{\mathcal{S}_{-i} := \{s_{j,k}\}_{j \neq i, k \in [M]}}$ be a realization of the prediction signals from all agents except i on all tasks. For a single task, we draw three random variables z_1, z_2, z_3. z_1 is a prediction signal uniformly randomly picked from all peer agents' prediction signals on that task. Excluding the picked signal z_1, we then a uniformly randomly pick a prediction signal and set it as z_2. Finally, we uniformly randomly pick a prediction signal as z_3, excluding both z_1 and z_2. z_1, z_2, z_3 are independent conditioned on the ground truth as agents' reports are conditional independent and we have infinite number of agents. Meanwhile, z_1 has the same error rates with the reference report z as they two come from the same random process. With infinite number of agents, z_2 and z_3 also have the same error rates as z. For the same reason to z (Proposition 3.1 and Assumption 4), z_1, z_2, z_3 each is i.i.d. across tasks. Therefore, with infinite tasks, we can know any statistics about z_1, z_2 and z_3 by counting corresponding frequencies on $\mathcal{S}_{-i}$. We can then establish the following three equations.

---

1. First-order equation: The first equation is based on the distribution 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:

$$\alpha_{-i} = p \cdot \Pr[z = 1 | y = 1] + (1 - p) \cdot \Pr[z = 1 | y = 0] = p \cdot (1 - e_z^+) + (1 - p) \cdot e_z^-. \tag{4}$$

2. Matching between two prediction signals: The second equation is derived from a second order statistics, namely the matching probability. We consider the matching-on-1 probability of two uniformly randomly picked prediction signals z_1, z_2 (on the same task, but from different peer agents). Denote this probability as $\underline{\beta_{-i} := \Pr[z_1 = 1, z_2 = 1]}$. This matching probability can be

Mechanism 2 SSR mechanism

1: For each task k, uniformly randomly pick at least 3 agents, assign task k to them, collect their reported predictions and generate the prediction signal for each prediction.
2: For each agent i and each task k she answers, uniformly randomly select one prediction signal $s_{j,k}$ from her peers' prediction signals on the same task and let the reference report $z_{i,k} := s_{j,k}$.
3: Solve Eqn.(4, 5, 6) to obtain e_z^-, e_z^+.
4: Pay each agent i for $q_{i,k}$ on task k by SSR_α if $e_{z_i}^+ + e_{z_i}^+ \neq 1$, and pay 0, otherwise.

written as a function of e_z^-, e_z^+:

$$\begin{aligned}\beta_{-i} &= p \cdot \Pr[z_1 = 1, z_2 = 1 | y = 1] + (1-p) \cdot \Pr[z_1 = 1, z_2 = 1 | y = 0] \\ &= p \cdot \Pr[z_1 = 1 | y = 1] \cdot \Pr[z_2 = 1 | y = 1] + (1-p) \cdot \Pr[z_1 = 1 | y = 0] \Pr[z_2 = 1 | y = 0] \\ &= p \cdot (1 - e_z^+)^2 + (1-p) \cdot (e_z^-)^2. \end{aligned} \tag{5}$$

3. Matching among three prediction signals: The third equation is obtained by going one order higher that, we check the matching-on-1 probability over three prediction signals z_1, z_2, z_3 drawn randomly from three different peer agents on the same task. Denote this probability as $\gamma_{-i} := \Pr[z_1 = z_2 = z_3 = 1]$. Similarly as Eqn. (5), we have:

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

Notice that all three parameters $\alpha_{-i}, \beta_{-i}, \gamma_{-i}$ can be perfectly estimated using $\mathcal{S}_{-i}$ with 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 5.3. *(p, e_z^-, e_z^+) are uniquely identified using Eqn.(4, 5, 6) under Assumption 3, that is, when $p \neq 0.5$ and the principal knows $\mathbb{1}(p > 0.5)$.*

The solution of Eqn.(4, 5, 6) can be expressed in closed form, which we present in Mechansim 3 in the finite sample setting. Now we have completed our mechanism. The full mechanism is presented in Mechanism 2. We further show that the three equations are both necessary and sufficient to estimate the error rates:

Theorem 5.4. *The higher order (≥ 4) matching equations do not bring in additional information.*

Theorem 5.3 shows that without ground truth data, knowing how frequently human agents reach consensus with each other will help us characterize their (average) subjective biases. Further, it implies that SSR mechanism is asymptotically (in M, N) preserving the information quantification as strictly proper scoring rules do and induces a strictly dominant strategy for agent to report truthfully, when z is informative (weakly dominant strategy otherwise). To see this, because both e_z^+, e_z^- are set to their true values, we have $\mathbb{E}[R(q_{i,k}, z)] = \mathbb{E}[S(q_{i,k}, y)]$. Formally,

Theorem 5.5. *When z is informative, asymptotically ($M, N \to \infty$) the expected score of SSR mechanism equals to the score of its corresponding strictly proper scoring rule S: $\mathbb{E}[R(q_{i,k}, z)] = \mathbb{E}[S(q_{i,k}, y)]$.*

Corollary 1. *SSR mechanism is dominantly truthful with infinite number of tasks and agents.*

Remark 1. *Theorem 5.3 and 5.5 rely on Proposition 3.1 and Assumptions 3 and 4. Proposition 3.1 and Assumption 4 guarantee that there exists, across the predictions of different tasks, a similar information pattern 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*

Mechanism 3 Estimation of e_z^+, e_z^-

1: Estimate $\widetilde{\alpha_{-i}}, \widetilde{\beta_{-i}}, \widetilde{\gamma_{-i}}$. Compute the following quantities:

$$a = \frac{\widetilde{\gamma_{-i}} - \widetilde{\alpha_{-i}}\widetilde{\beta_{-i}}}{\widetilde{\beta_{-i}} - (\widetilde{\alpha_{-i}})^2},\ b = \frac{\widetilde{\alpha_{-i}}\widetilde{\gamma_{-i}} - (\widetilde{\beta_{-i}})^2}{\widetilde{\beta_{-i}} - (\widetilde{\alpha_{-i}})^2},\ \underline{x} = \frac{a - \sqrt{a^2 - 4b}}{2},\ \overline{x} = \frac{a + \sqrt{a^2 - 4b}}{2}$$

2: Denote by $\underline{e}, \overline{e}$ as the $\underline{x}, \overline{x}$ that are closer and further to $\widetilde{\alpha_{-i}}$ respectively:

$$\underline{e} = \text{argmin}_{x \in \{\underline{x}, \overline{x}\}} |x - \widetilde{\alpha_{-i}}|, \quad \overline{e} = \text{argmax}_{x \in \{\underline{x}, \overline{x}\}} |x - \widetilde{\alpha_{-i}}|$$

3: If $p < 0.5$: $\widetilde{e_z^-} := \underline{e}$, $\widetilde{e_z^+} := 1 - \overline{e}$; else if $p > 0.5$: $\widetilde{e_z^-} := \overline{e}$, $\widetilde{e_z^+} := 1 - \underline{e}$.

distinguish the case when agents are truthfully reporting from the case that agents are misreporting by reverting their observations. This is because for every possible tuple (p, e_z^-, e_z^+) resulted by truthful reporting from agents, consider the following counterfactual world: relabeling $0 \to 1$ and $1 \to 0$, we will have another distribution of observations characterized by the tuple $(1 - p, e_z^+, e_z^-)$. Then agents misreporting will lead to a distribution with parameters being the same as (p, e_z^-, e_z^+). Thus the mechanism designer cannot tell the above two cases apart. Some work [14] relaxes Assumption 3 by excluding the "relabeling equilibrium" from consideration.

We will show in the next section, SSR mechanism is also dominantly truthful with 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 prediction signals of her peer agents' reports $\mathcal{S}_{-i}$: z will be decided by agents $j \neq i$'s reports $\mathcal{S}_{-i}$. $R(\cdot)$ not only has z as input, but its definition also depends on e_z^+ and e_z^-, which will be learned from $\mathcal{S}_{-i}$. (2) When making decisions on reporting, we show under our mechanisms agents can choose to be oblivious of how much error presents in others' reports. This removes the practical concern of implementing a particular Nash Equilibrium. (3) Another salient feature of our mechanism is that we have migrated the cognitive load for having prior knowledge from agents to the mechanism designer. Yet we do not assume the designer has direct knowledge neither; instead we will leverage the power of estimation from reported data to achieve our goal.

5.3 Finite sample analysis

With finite M, N, there are multiple reasons that we won't be able to obtain perfect estimates of e_z^+, e_z^-. For instance, in forming Eqn.(4, 5, 6), the error rates of two randomly picked prediction signals z_2, z_3 will not have the exactly same error rates with z. However when the number of agent is large enough, we will show that the error rates of z_2, z_3 can approximate these e_z^+, e_z^- with small and diminishing errors (as a function of number of agents N). This can factor into the errors in estimating β_{-i}. Furthermore, the algorithm's estimates of the following three parameters for each agent i, $\alpha_{-i}, \beta_{-i}, \gamma_{-i}$, are not perfect.

All three parameters $\alpha_{-i}, \beta_{-i}, \gamma_{-i}$ can be estimated from agents' reports, without the need of knowing any ground truth labels. Let k_1, k_2, k_3 be the three agents whose prediction signals are selected as z_1, z_2, z_3 for each task $k \in [M]$ (In practice, we only need to assign task k to these three randomly selected agents). Then we estimate:

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

We then solve the system of equations (4, 5, 6) with these estimates to obtain estimated error rates e_z^+, e_z^-. We present the solution in Mechanism 3.

We give a statistical consistency analysis for this estimation procedure for this finite sample setting. We bound the estimation error in estimating reports' error rate as a function of M and N. The first source of errors is due to the imperfect estimations of $\beta_{-i}, \gamma_{-i}, \alpha_{-i}$. The second one is due to estimation errors for matching probability with heterogeneous agents. Formally we have the following theorem:

LEMMA 5.6. *$\widetilde{e_z^+}, \widetilde{e_z^-}$ given by Mechanism 3 satisfy $|\widetilde{e_z^+} - e_z^+| \le \epsilon$, $|\widetilde{e_z^-} - \widetilde{e_z^-}| \le \epsilon$ with probability at least $1-\delta$, where $\epsilon := O\big(\frac{1}{N} + \sqrt{\frac{\ln \frac{1}{\delta}}{M}}\big)$, which can be made arbitrarily small with increasing M and N.*

Denote by $\Delta := (1-p)(1-e_z^- - e_z^+)$. The above estimation of e_z^+, e_z^- further leads to the following the above consistency result:

THEOREM 5.7. *For the scoring function $\widetilde{R}(\cdot)$ defined for SSR_α using $\widetilde{e_z^+}, \widetilde{e_z^-}$, when M, N are large enough s.t. $\epsilon \le (1 - e_z^- - e_z^+)/4$, with probability at least $1-\delta$,*

$$|\widetilde{R}(q_i, z) - R(q_i, z)| \le \frac{12\epsilon \cdot \max S}{\Delta^2},\ \forall q_i \in [0,1], z \in \{0,1\},$$

where $\max S$ is the maximum score of the underlying SPSR that $\widetilde{R}$ builds on. This further implies that

$$|\mathbb{E}[\widetilde{R}(q_i, z)] - \mathbb{E}[R(q_i, z)]| \le \frac{12\epsilon \cdot \max S}{\Delta^2},\ |\mathbb{E}[\widetilde{R}(q_i, z)] - \mathbb{E}[S(q_i, y)]| \le \frac{12\epsilon \cdot \max S}{\Delta^2},\ \forall q_i \in [0,1]$$

Now we present the incentive guarantees in finite sample regime under noisy estimations. We first note that any linear transformation of a particular SSR mechanism preserves its incentive property. To simply our analysis, we will first perform the following operation to "cancel" the effects of noisy estimation of e_z^+, e_z^- in the denominator of $R(\cdot)$: $\widetilde{R}(q_i, z) := (1 - \widetilde{e_z^+} - \widetilde{e_z^-}) \cdot \widetilde{R}(q_i, z)$ - note the above linear transform (independent of agent's reports) does not change the incentive property of SSR.

THEOREM 5.8. *When z is informative, set M, N large enough but finite, SSR mechanism returns a score that is $\epsilon(M, N)$ close to the score of its corresponding strictly proper scoring rules, where $\epsilon(M, N) = O\big(\frac{1}{N} + \sqrt{\frac{\ln M}{M}}\big)$ is a diminishing term in both M and N. Further, for each agent i, it is a strictly dominant strategy to truthfully report $q_{i,k}, \forall k$ when $S(q, y)$ is strongly concave and Lipschitz in q for any $y \in \{0,1\}$ and M, N are sufficiently large.*

The intuition about dominant truthfulness part is that when M, N are sufficiently large, the estimation error is too small such that the deviation gain through utilizing the error cannot surpass the loss in the true score, and the qualified M, N are determined by the curvature of $S(\cdot)$.

COROLLARY 2. *When SPSR $S(q, y)$ is strongly concave and Lipschitz in q for all $y \in \{0,1\}$, the SSR mechanism built upon $S(\cdot)$ is dominantly truthful with finite but sufficiently large N and M.*

For example, Log scoring rule over interval $[0.01, 0.99]$ is strongly concave and Lipschitz.[5]

6 EMPIRICAL STUDIES

Using 14 real-world human forecasting datasets, we demonstrate that without the need of accessing ground truth, SSR mechanism demonstrate stronger correlation with the true scores given by SPSR (which use ground truth outcome) than the other peer prediction methods across different datasets we tested over.

[5]When log scoring rule is applied, the range of the prediction is usually restricted to a closed interval excluding point 0 and 1, e.g., [0.01, 0.99]. This is because log scoring rule is not well-defined (infinite) when the prediction is 0 (or 1) while the ground truth is 1 (or 0).

Items	G1	G2	G3	G4	H1	H2	H3	M1a	M1b	M1c	M2	M3	M4a	M4b
# of questions (original)	94	111	122	94	88	88	88	50	50	50	80	80	90	90
# of agents (orginal)	1972	1238	1565	7019	768	678	497	51	32	33	39	25	20	20
After applying the filter														
# of questions	94	111	122	94	72	80	86	50	50	50	80	80	90	90
# of agents	1409	948	1033	3086	484	551	87	51	32	33	39	25	20	20
Avg. # of answers per question	851	533	369	1301	188	252	33	51	32	33	39	18	20	20
Avg. # of answers per agent	57	62	44	40	28	37	33	50	50	50	80	60	90	90
Majority vote correct ratio (%)	0.90	0.92	0.95	0.96	0.88	0.86	0.92	0.58	0.76	0.74	0.61	0.68	0.62	0.72

Table 1. Statistics about binary-outcome datasets from GJP, HFC and MIT datasets

6.1 Setting

We evaluate the properties of SSR mechanism (built upon three popular SPSR) with 14 real-world forecasting datasets and compare the results to those of other four popular existing peer prediction methods. In what follows, we introduce the details of these settings.

6.1.1 Datasets. We conduct our experiments on 14 datasets from three human forecasting and crowdscourcing projects: the Good judgment Project (GJP), the Hybrid Forecasting Project (HFC) and an MIT collected human judgment datasets. These three projects are different in both the populations of participants, forecast topics and elicitation methods.

GJP datasets [2]. It contains four datasets on geopolitical forecasting questions. The four datasets, denoted by G1~G4, was collected from 2011 to 2014 respectively. They have different forecasting questions and forecasters. Each forecaster has a single probabilistic prediction for a question she answered in the datasets.

HFC datasets [12]. It contains three datasets, denoted by H1~H3, collected from the Hybrid Forecast Competition organized by IARPA in 2018. The three datasets share the same forecasting questions about geopolitics, finance, economics, etc, but have different forecasters and collecting methods. These three datasets record multiple probabilistic predictions each forecaster made at different dates. We used the final prediction made by a forecaster on a question she answered.

MIT datasets [25]. It contains seven datasets, denoted as M1a, M1b, M1c, M2, M3, M4a, M4b, with different questions and forecasters. The questions ranges from the capital of states to the price interval that artworks belong to, to some trivia questions. The forecasters were students in class and colleagues in labs. In datasets M1a, M1b, M4a, M4b, forecasters made binary vote on a forecasting question. In datasets M1c, M2, M3, forecasters gave a probabilistic prediction.

We focus on the forecasting questions with binary outcomes in these datasets. We filtered out the questions with less than 10 submitted predictions and the participants who predicted on less than 15 questions. No questions were filtered out from GJP and MIT datasets and only a few from HFC datasets. Basic statistics of these datasets are presented in Table 1.

6.1.2 SPSR. We consider three SPSR: Brier score, log scoring rule, and rank-sum scoring rule. The first two are the most widely adopted scoring rules, and they are equivalent to squared error and cross-entropy loss, respectively, for measuring the accuracy of predictions. The rank-sum scoring rule can be written as an affine transformation (depending on the number of tasks in each ground truth category) of AUC-ROC metric, [23]. Therefore, it is also of interest to us.

In the experiments, we adopt the convention used the in the GJP for Brier score that it ranges from 0 to 2 and a smaller score corresponds to a higher accuracy.[6] To align with Brier score, we also use a log scoring rule and a rank-sum score rule that a smaller score corresponds to a higher accuracy and the minimum possible score is 0.

Let $[M_i]$ be the set of tasks answered by agent i. Recall that $q_{i,k}$ and y_k are agent i's prediction and the ground truth for task k, respectively. The exact formulas for the three scoring rules we used are as follows:

- **Brier score**: $S^{\text{Brier}}(q_{i,k}, y_k) = (q_{i,k} - y_k)^2 + \left((1 - q_{i,k}) - (1 - y_k)\right)^2 = 2(q_{i,k} - y_k)^2$.
  An agent's accuracy score under Brier score is the mean Brier score $\frac{1}{M_i} \sum_{k \in [M_i]} S^{\text{Brier}}(q_{i,k}, y_k)$.
- **Log scoring rule**: $S^{\log}(q_{i,k}, y_k) = \log(q_{i,k})$ if $y_k = 1$; and $S^{\log}(q_{i,k}, y_k) = \log(1 - q_{i,k})$ if $y_k = 0$.
  An agent's accuracy under log scoring rule is also the mean score $\frac{1}{M_i} \sum_{k \in [M_i]} S^{\log}(q_{i,k}, y_k)$.
  As it is unbounded in the worst case, we change all predictions with value 1 to 0.99 and predictions with value 0 to 0.01 to ensure a well-defined score.
- **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 $\psi\left(q_{i,k} | \{q_{i,k'}\}_{k' \in [M_i]}\right) := \sum_{k' \in [M_i]} \mathbb{1}(q_{i,k'} < q_{i,k}) - \sum_{k' \in [M_i]} \mathbb{1}(q_{i,k'} > q_{i,k})$ is the rank of prediction $q_{i,k}$ in all agent i's predictions. Then, agent i's rank-sum score S_i^{rank} is defined: $S_i^{\text{rank}} = \sum_{k \in [M_i]} S^{\text{rank}}(q_{i,k}, y_k)$.[7] The range of the score increases with the number of answered tasks quadratically. We normalize the score using $1 + \frac{4}{M_i^2} S_i^{\text{rank}}$ to range $[0, 2]$.

6.1.3 Treatments. Though existing peer prediction methods are not designed for recovery of SPSR, we add comparisons to them for completeness of our study.[8] In particular, we'd like to understand whether in practice SSR has the advantage of revealing the true scores given by SPSR while not accessing ground truth information.

In our experiments, we consider four popular existing peer prediction methods, serving as comparisons to SSR: proxy scoring rule (PSR) with extremized mean [34], peer truth serum (PTS) [27], correlated agreement (CA) [31], determinant mutual information (DMI) [14].

PSR is to directly apply the SPSR w.r.t. an unbiased proxy of the ground truth, and Witkowski et al recommended using the extremized mean of the reported predictions as the unbiased proxy, when there is no verification data available [34]. Using different SPSR as the building block, we can get different PSR. PTS, CA, DMI are not build upon SPSR and are designed to elicit a categorical label instead of a probabilistic prediction. When applied them on datasets with probabilistic predictions, we assume that a categorical label is drawn from the probabilistic prediction and we compute an asymptotically consistent estimator of their expected scores, where the expectation is taken over the drawn of the categorical label.

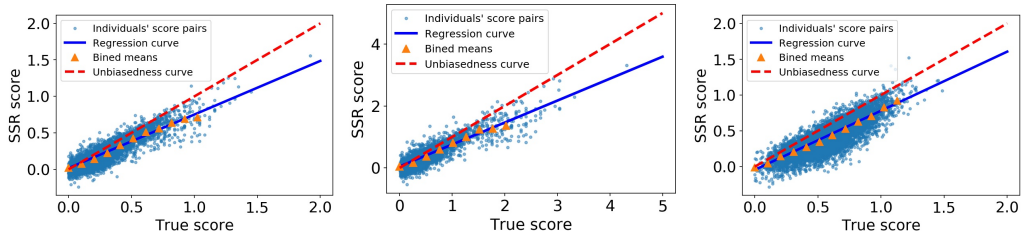
6.2 Main results

Unbiasedness of SSR. We exam to what extend SSR recover the true accuracy scores given by different SPSR. We compute the true mean score and mean SSR score of each human forecaster in all datasets.

[6] This is different from using SPSR as a payment method, where the higher the better. We can transfer between these two usages by applying a negative scalar.

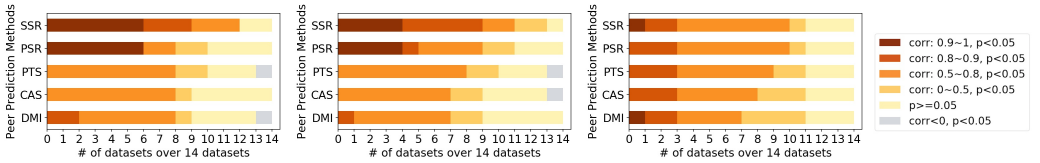
[7] The AUC-ROC of agent i is equal to $\frac{1}{2}\left(1 - \frac{1}{M_i^+(M_i - M_i^+)} S_i^{\text{rank}}\right)$, where $M_i+ := \sum_{k' \in [M_i]} \mathbb{1}(y_{k'} = 1)$ [23].

[8] We do not intend to claim 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.



(a) Brier ($y = 0.787 \cdot x + 0.001$) (b) Log ($y = 0.790 \cdot x - 0.005$) (c) Rank-sum ($y = 0.839 \cdot x - 0.057$)

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



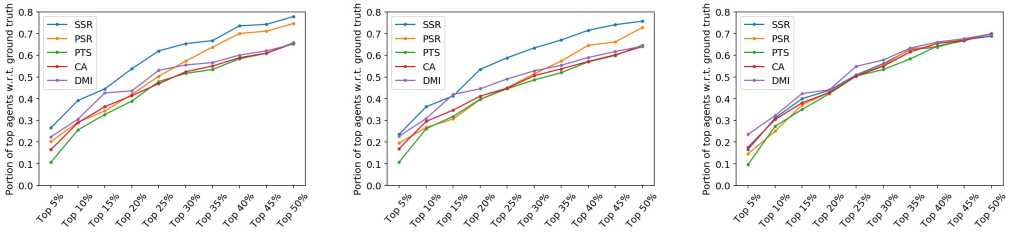
(a) Brier score (b) Log scoring rule (c) Rank-sum scoring rule

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

The pairs of true mean accuracy score and mean SSR score of every individual in the 14 datasets are illustrated by blue dots in Fig 1. It is clear that most of them concentrate around $y = x$, which demonstrates the unbiasedness of SSR scores. Then, we separate forecasters into different bins w.r.t. their true scores. For Brier score and rank-sum scoring rule, the centers of the bins are from 0 to 2.0 with a width of 0.05. For log scoring rule, the centers of the bins are from 0 to 5 with a width of 0.1. For forecasters in each bin, we then calculate the mean SSR score of these forecasters (we ignore bins with less than 20 forecasters). We find that for users at same true score level, their SSR scores are also at at similar level. These are illustrated by orange triangles in Fig 1. Finally, we draw the linear regression curves on these binned means such that each true accuracy level is weighted uniformly in the regression (blue curve in Fig 1). The slope for the three curves are all around 0.8, while the intercepts are all round 0. This shows that the average SSR score is extremely close to the true accuracy score when the true accuracy score is small. In other words, SSR can calibrate the true accuracy almost perfectly for sophisticated forecasters. Given most agents have a true accuracy score better than uniformly randomly guessing 0 and 1 (which is 1 in Brier score and rank-sum score and 2.3 in log score) in these 14 datasets, SSR approximate the true scores well for most of the time.

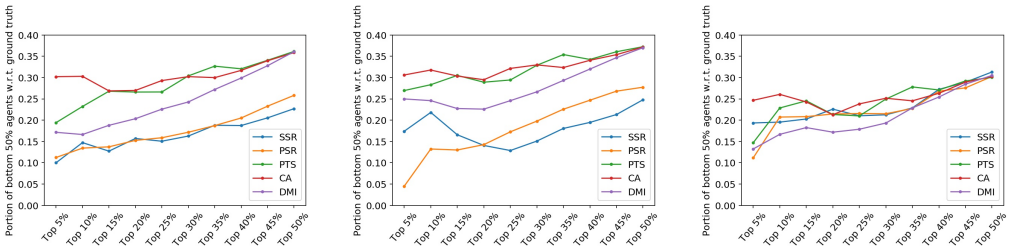
Correlation with SPSR. We exam the correlations between agents' peer prediction scores and true accuracy scores given by the three SPSR, Brier score, log scoring rule and rank-sum scoring rule. When a SPSR is chosen as the true score, we also use this SPSR as the underlying scoring rule called by SSR and PSR. PTS, CA and DMI scores are independent from which SPSR is used. We adjust the scores such that a lower score corresponds to a higher accuracy (or a higher payment to the agents) in the context of each peer prediction method.

We exam these correlations on each dataset independently, and categorize the level of correlations according to the Pearson's correlation coefficient and p-values. As shown in Fig 2, we find that for Brier score, and log scoring rule, SSR achieves a Pearson's correlation COEFFICIENT > 0.8 on 9 out of 14 datasets. The second best, PSR, achieves a COEFFICIENT > 0.8 on at most 6 out of 14 datasets. PTS and CAS do not have a COEFFICIENT > 0.8 on any datasets, while DMI achieves



(a) Mean squared loss (b) Mean cross-entropy loss (c) AUC-ROC

Fig. 3. The portion of top $t\%$ forecasters w.r.t. 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) Mean squared loss (b) Mean cross-entropy loss (c) AUC-ROC

Fig. 4. The portion of bottom 50% forecasters w.r.t. 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).

COEFFICIENT > 0.8 on at most 2 of the datasets. For rank-sum scoring rule, all peer prediction scores achieve similar levels of correlation among 14 datasets, while SSR are better than the others. We observe similar results on Spearman's correlation test (Fig 5 in the Appendix [18]). This result on Spearman's (rank) test, in particular, implies that SSR mechanism rank the agents in a similar order of agents' true expertise.

Expert identification. We exam to what extent different peer prediction scores can identify top performing experts. We rank the forecasters according to one of three most-widely used loss function (mean squared loss, mean cross-entropy loss, and AUC-ROC). We focus on two metrics about expert identification: i. percent of true top $t\%$ forecasters in the top $t\%$ forecasters selected by a peer prediction methods, ii. percent of below-average forecasters, the bottom 50% forecasters, in the top $t\%$ forecasters selected by a peer prediction methods. Results are shown in Fig 3 and Fig 4. We find that for both mean squared loss and mean cross-entropy loss, in the top $t\%$ forecaster selected by SSR, there are more true top $t\%$ forecasters, than in the top forecasters selected by other peer prediction scores for $t\%$ ranges from 5% to 50%. Meanwhile, there are less below-average forecasters in the top $t\%$ forecasters top $t\%$ by SSR and PSR than by the other peer prediction scores. For AUC-ROC, different peer prediction scores have similar performance, while SSR and DMI are slightly better than the others. These results echo the results about the correlation of peer prediction scores w.r.t. different SPSR.

7 CONCLUDING REMARKS

We propose SSR to quantify the value of elicited information in IEWV settings, as strictly proper scoring rules do for the *with* verification setting. SSR also induce truthful reporting in strictly dominant strategy for eliciting probabilistic predictions. SSR contribute to both the SPSR and peer

prediction literature. Our findings are both verified analytically and empirically. Our work opens up the study of calibrating the value of information for the peer prediction setting.

ACKNOWLEDGMENTS

This research is based upon work supported in part by National Science Foundation (NSF) under Grant No. CCF-1718549, 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. 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.

REFERENCES

[1] Dana Angluin and Philip Laird. 1988. Learning from noisy examples. *Machine Learning* 2, 4 (1988), 343–370.
[2] 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.
[3] Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthey Weather Review* 78, 1 (1950), 1–3.
[4] Tom Bylander. 1994. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the seventh annual conference on Computational learning theory*. ACM, 340–347.
[5] 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.
[6] Luca De Alfaro, Michael Shavlovsky, and Vassilis Polychronopoulos. 2016. Incentives for truthful peer grading. *arXiv preprint arXiv:1604.03178* (2016).
[7] Alexander Frankel and Emir Kamenica. 2019. Quantifying information and uncertainty. *American Economic Review* 109, 10 (2019), 3650–80.
[8] 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.
[9] Alice Gao, James R Wright, and Kevin Leyton-Brown. 2016. Incentivizing evaluation via limited access to ground truth: Peer-prediction makes things worse. *arXiv preprint arXiv:1606.07042* (2016).
[10] Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102, 477 (2007), 359–378.
[11] Naman Goel and Boi Faltings. 2018. Deep Bayesian Trust : A Dominant and Fair Incentive Mechanism for Crowd. arXiv:cs.GT/1804.05560
[12] IARPA. 2019. Hybrid Forecasting Competition. https://www.iarpa.gov/index.php/research-programs/hfc?id=661.
[13] 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.
[14] Yuqing Kong. 2020. Dominantly Truthful Multi-task Peer Prediction with a Constant Number of Tasks. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2398–2411.
[15] Yuqing Kong and Grant Schoenebeck. 2016. Equilibrium selection in information elicitation without verification via information monotonicity. *arXiv preprint arXiv:1603.07751* (2016).
[16] 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.
[17] Yuqing Kong and Grant Schoenebeck. 2019. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. *ACM Transactions on Economics and Computation (TEAC)* 7, 1 (2019), 2.
[18] Yang Liu, Juntao Wang, and Yiling Chen. 2018. Surrogate scoring rules. *arXiv preprint arXiv:1802.09158* (2018).
[19] 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.
[20] 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.

[21] Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science* 51, 9 (2005), 1359 –1373.

[22] 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.

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

[24] Dražen Prelec. 2004. A Bayesian Truth Serum for Subjective Data. *Science* 306, 5695 (2004), 462–466.

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

[26] 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)*.

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

[28] Leonard J. Savage. 1971. Elicitation of Personal Probabilities and Expectations. *J. Amer. Statist. Assoc.* 66, 336 (1971), 783–801.

[29] Clayton Scott. 2015. A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels.. In *AISTATS*.

[30] 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.

[31] 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.

[32] Brendan van Rooyen and Robert C Williamson. 2015. Learning in the Presence of Corruption. *arXiv preprint:1504.00091* (2015).

[33] Robert L. Winkler. 1969. Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.* 64, 327 (1969), 1073–1078.

[34] Jens Witkowski, Pavel Atanasov, Lyle H Ungar, and Andreas Krause. 2017. Proper proxy scoring rules. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[35] Jens Witkowski, Yoram Bachrach, Peter Key, and David C. Parkes. 2013. Dwelling on the Negative: Incentivizing Effort in Peer Prediction. In *HCOMP'13*.

[36] 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)*.