Surprisingly Popular Voting Recovers Rankings, Surprisingly!*

Hadi Hosseini
Pennsylvania State University
hadi@psu.edu

Debmalya Mandal
University of Warwick
Debmalya.Mandal@warwick.ac.uk

Nisarg Shah
University of Toronto
nisarg@cs.toronto.edu

Kevin Shi
University of Toronto
kevins.shi@mail.utoronto.ca

Abstract

Classical democratic approaches for aggregating individual *votes* only work when the opinion of the majority of the crowd is relatively accurate. A clever recent approach, *surprisingly popular algorithm*, elicits additional information from the individuals, namely their *prediction* of other individuals' votes, and provably recovers the ground truth even when experts are in minority. This approach works well when the goal is to pick the correct option from a small list, but when the goal is to recover a true ranking of the alternatives, a direct application of the approach requires eliciting too much information. We explore practical techniques for extending the surprisingly popular algorithm to ranked voting given only partial votes and predictions, and design robust aggregation rules to recover true rankings. In particular, we introduce six elicitation formats with varying information requirements to facilitate ranked versions of surprisingly popular algorithm. Through a crowdsourcing experiment on MTurk, we demonstrate that even a little prediction information helps surprisingly popular voting outperform classical approaches across a diverse set of domains with varying difficulty levels.

1 Introduction

The wisdom of the crowd has been the default choice for uncovering the ground truth. Suppose we wish to determine the true answer to the question: "Is Philadelphia the capital of Pennsylvania?" Condorcet's Jury Theorem suggests that if we elicit votes from a large crowd, the majority answer will be correct with high probability even if, on average, the crowd is only slightly more accurate than a random selection. However, in some domains the crowd can be highly inaccurate and experts may be in minority. For example, when the very question listed above is posed to real crowds, the majority answer is often (the incorrect) 'yes' [12].

To circumvent this difficulty and uncover the ground truth even when the majority is wrong, Prelec et al. [31] introduce the *surprisingly popular* (SP) algorithm. This algorithm asks each individual not only what she thinks the answer is (the *vote*), but also what fraction of the other

*This preprint has been accepted to the Journal of Artificial Intelligence Research (JAIR). It extends a preliminary version published in the Proceedings of 30th International Joint Conference on Artificial Intelligence (IJCAI), pp. 245-251, 2021.

participants she thinks will say yes/no (the *prediction*). Then, instead of simply selecting the majority (i.e. popular) answer, the algorithm selects the answer that is *surprisingly popular*, i.e., whose actual frequency in the votes is greater than its average predicted frequency. They show that as the crowd gets larger in the limit, this approach will provably recover the correct answer with probability 1, even if the crowd is less accurate than a random selection on average.

The intuition behind their algorithm, borrowed from their work, is as follows. Suppose there are two hypothetical worlds, one where Philadelphia is the capital and one where it is not. In the former world, a greater fraction (say 90%) would say 'yes' than the fraction (say 60%) that would say 'yes' in the latter. However, the 60% of the people who believe the correct world is the former would predict the frequency of 'yes' to be 90%, whereas the remaining 40% would predict it to be 60%. This would make the average predicted frequency of 'yes' to be somewhere between 60% and 90%, higher than its actual frequency of 60%. In other words, the majority but incorrect answer 'yes' would be surprisingly *unpopular* while 'no' would be surprisingly popular and correct.

Several works have demonstrated the effectiveness of this approach in a wide range of domains [23, 25, 28, 31, 34, 37]. Prediction questions have also been used to boost the accuracy of surveys on social networks [15]. Prelec et al. [31] show how to apply their approach to questions with non-binary votes and non-binary ground truth. When the true answer lurks among r options, their approach requires each individual to predict the exact frequency of each of r options among other individuals' votes. We are interested in ranked voting, i.e., when the ground truth is a ranking of m alternatives. Note that in this case, the approach of Prelec et al. [31], which we refer to as *surprisingly popular (SP) voting*, would require eliciting predictions in the form of a distribution over $r = m!$ options, which is clearly infeasible for even moderate values of m. Thus, the main research questions we address are:

> *How do we extend surprisingly popular voting to effectively recover a ground truth ranking of alternatives? If we elicit partial votes and predictions, how do we aggregate them and what information-accuracy tradeoff does this offer?*

1.1 Our Contributions

We focus on eliciting only *ordinal* vote and prediction information. For the *vote*, we ask individuals to provide their opinion of either just the top alternative of the ground truth ranking (*Top*) or the full ground truth ranking (*Rank*). For the *prediction*, informally, we ask individuals to predict either just a single alternative (*Top*) or a ranking of alternatives (*Rank*) based on the other individuals' votes. The exact prediction elicited under various conditions is described in Section 3. In addition to these four elicitation formats, we use as benchmark two classical elicitation formats in which Top and Rank votes are elicited but no prediction is elicited. Because the SP algorithm of Prelec et al. [31] does not work on partial votes and predictions, we first design a novel aggregation method for such partial information.

Next, we conduct an empirical study with 720 participants from Amazon's Mechanical Turk platform. We ask the participants questions on geography, movies, and artwork which admit a ground truth ranking of four alternatives and elicit their responses in the aforementioned six elicitation formats. We compare the different elicitation formats using four metrics: difficulty (measured through response time as well as perceived difficulty), expressiveness, error in recovering the ground truth top alternative, and error in recovering the ground truth ranking.

Our results show that even when the vote and prediction information are individually no better than random guesses, by combining the two pieces of information SP voting performs significantly better. Further, it outperforms a whole slew of conventional voting rules which ignore prediction information and only aggregate the votes. We also observe that when it is necessary to choose between eliciting more complex vote information and eliciting more complex prediction information, the latter may be the right choice.[1]

1.2 Related Work

Our work builds on the surprisingly popular algorithm of Prelec et al. [31]. This approach in turn builds on its precursor, the Bayesian truth serum (BTS) [30], which also uses participants' predictions, but for a different objective: to decide payoffs to the participants which incentivize them to honestly report their votes and predictions. Recently Chen et al. [8] proposed an algorithm to recover the true state of the world, however, they ask users to report direct predictions over the ground truth, and for more than two states, their required condition is different than [31].

Prediction markets [4, 9], and quadratic voting [22] are alternative approaches that solicit additional information from the voters other than their preferences. The additional information lets the mechanisms select alternatives that could be different than the ones selected by voting rule from the preferences. For example, prediction markets ask participants to place a bet on their vote while quadratic voting asks voters to express the degree of their preferences by purchasing additional votes on a given matter. Note that, on these platforms, it is indeed possible for the minority experts to override the majority preferences through the additional information e.g. purchasing additional votes in quadratic voting. However, these approaches mainly focus on the design of payments to guarantee that truth-telling forms an equilibrium, and in order to recover ground truth at an equilibrium, we need to incentivize the minority experts, which might not be possible for all questions.

Our work is also closely related to the literature on peer prediction [27, 30] which attempts to elicit information from strategic agents when verification is not possible. Classical peer prediction mechanisms are non-minimal in the sense that they also use participants' predictions of other participants' signals. A flurry of work in the last decade [2, 10, 21] has developed improved peer prediction mechanisms often with better equilibrium guarantees [20, 36]. See Faltings and Radanovic [14] for a survey.

In the context of voting, Schoenebeck and Tao [35] recently proposed the wisdom of crowd voting, an adaptation for the surprisingly popular algorithm. They consider a similar setup as ours where the voters receive private signals conditioned on the true underlying choice, according to a common probabilistic model. However, their goal is to output the alternative preferred by the majority of voters in a strong Bayesian Nash equilibrium. Moreover, they consider the setting of two alternatives, whereas our main contribution lies in generalizing the surprisingly popular algorithm to ranked choices over a set of alternatives.

These recent approaches stand in contrast to a large body of work on epistemic social choice [29] and noisy voting [7], which build on the seminal work of de Condorcet [13], Galton [16], and Young [38]. Some of this literature focuses on statistical models of errors in participants' votes such as the Mallows model, the Bradley-Terry model, the Thurstone-Mosteller model, and the Plackett-

[1] All relevant datasets and codes are publicly available at https://github.com/debmandal/Surprisingly-Popular-Voting.

Luce model. However, all these models assume that a participant is ever-so-slightly more likely to report the correct option than an incorrect option. Hence, approaches based on these models can fail to recover the ground truth when the majority of the crowd is misinformed.

Besides epistemic social choice, our work is also related to the literature on truth discovery through crowdsourcing [33, 39]. Building on the seminal work of [11] these works usually adopt Expectation-Maximization (EM) algorithm to estimate the error rates of the workers and then use these error estimates on a new task. However, approaches based on these methods often require lots of samples and don't work when the error rates of the workers change from task to task.

Finally, our work is reminiscent of a recent flurry of work on the elicitation-distortion trade-off in computational social choice [1, 3, 19, 24, 26]. In this line of work, there is no ground truth; instead, participants have subjective preferences and the goal is to identify the decision that maximizes the social welfare. Rather than directly eliciting participants' utility functions, various elicitation formats are used to elicit partial preferences to analyze the trade-off between the amount of information elicited and the approximation to social welfare (called distortion). Our setting replaces the distortion with its counterpart, that is, the accuracy of recovering an underlying ground truth. Moreover, our methods ask voters to report their votes and prediction about other voters' preferences. However, such a prediction report is a full distribution over all preferences and might be too prohibitive to elicit. So we elicit partial prediction reports from the voters – most likely top alternative and most likely ranking. We study trade-offs between various types of prediction reports and accuracy, and observe that moving from top prediction report, to rank prediction report increases communication from the voters, but improves accuracy i.e. average distance from the ground truth ranking.

2 Model

Let A be a set of m alternatives and $\mathcal{L}(A)$ be the set of rankings over A. For a ranking $\sigma \in \mathcal{L}(A)$ and $x \in \{1, \ldots, m\}$, let $\sigma(x)$ be the alternative in the x^{th} highest position in σ.

SP voting uses a Bayesian model; in the following, we present a special case of the model for ranked voting. There exists a ground truth ranking $\pi^* \in \mathcal{L}(A)$ drawn from a *prior* $\mathcal{P}$. There are n voters; each voter i observes a noisy ranking $\sigma_i \in \mathcal{L}(A)$ drawn from a *signal distribution* $\Pr_s(\cdot|\pi^*)$. The voters know both the prior $\mathcal{P}$ and the signal distribution $\Pr_s(\cdot|\pi^*)$; however, the principal is unaware of both. Following Prelec et al. [31], we assume that $\mathcal{P}(\pi), \Pr_s(\sigma|\pi) > 0$ for all rankings $\sigma, \pi \in \mathcal{L}(A)$ to avoid degeneracy.

Conventional voting would ask each voter i to simply report her observed noisy ranking σ_i and use a voting rule such as the Kemeny rule or Borda count to aggregate the reported rankings. SP voting additionally asks each voter i to make inferences about the reports of other voters. Given her observed noisy ranking σ_i and the prior $\mathcal{P}$, voter i can compute a posterior distribution over the ground truth, given by

$$\Pr_g(\pi^*|\sigma_i) = \frac{\Pr_s(\sigma_i|\pi^*) \cdot \mathcal{P}(\pi^*)}{\sum_{\pi' \in \mathcal{L}(A)} \Pr_s(\sigma_i|\pi') \cdot \mathcal{P}(\pi')}. \tag{1}$$

In turn, the voter can also infer a distribution over the noisy ranking σ_j observed by another voter j:

$$\Pr_o(\sigma_j|\sigma_i) = \sum_{\pi^* \in \mathcal{L}(A)} \Pr_s(\sigma_j|\pi^*) \cdot \Pr_g(\pi^*|\sigma_i).$$

SP voting asks each voter i to report not only her observed noisy ranking σ_i (the *vote*), but also her inferred distribution $\Pr_o(\cdot|\sigma_i)$ over other voters' noisy rankings (the *prediction*). Given these reports, for a ranking $\pi \in \mathcal{L}(A)$, let $f(\pi) = 1/n \cdot \sum_{i=1}^{n} \mathbb{1}[\sigma_i = \pi]$ denote the fraction of voters who vote π and $g(\cdot|\pi)$ denote the average of reported predictions $\Pr_o(\cdot|\sigma_i)$ across all voters i with $\sigma_i = \pi$. Then, the SP algorithm of Prelec et al. [31] computes the prediction-normalized vote count for each possible ground truth π as

$$\bar{V}(\pi) = f(\pi) \cdot \sum_{\pi' \in \mathcal{L}(A)} \frac{g(\pi'|\pi)}{g(\pi|\pi')}. \tag{2}$$

and selects the ranking π with largest prediction-normalized score.

Let us illustrate this algorithm through an example with two alternatives a and b. Here the goal is to determine whether the true ranking is $a \succ b$ or $b \succ a$. The prior over these two orders is

$$\mathcal{P}(a \succ b) = 0.4 \text{ and } \mathcal{P}(b \succ a) = 0.6.$$

If the ground truth is $a \succ b$ then the signals are distributed as follows.

$$\Pr_s(a \succ b \mid a \succ b) = 0.4 \text{ and } \Pr_s(b \succ a \mid a \succ b) = 0.6$$

On the other hand, if the ground truth is $b \succ a$ then the signals are distributed as follows.

$$\Pr_s(a \succ b \mid b \succ a) = 0.3 \text{ and } \Pr_s(b \succ a \mid b \succ a) = 0.7$$

Now suppose that the true ranking is $a \succ b$ which means that the votes (or signals) are distributed as $f(a \succ b) = 0.4$ and $f(b \succ a) = 0.6$. In this case, the majority will pick the ranking $b \succ a$ which is incorrect. We also assume noiseless setting i.e. $g(\pi' \mid \pi) = \Pr_o(\pi' \mid \pi)$ for any π, π'.

Now we can use Equation (1) to compute the posterior distribution over the ground truth rankings.

$$\Pr_g(a \succ b \mid a \succ b) = 0.471, \Pr_g(b \succ a \mid a \succ b) = 0.529$$

and

$$\Pr_g(b \succ a \mid b \succ a) = 0.364, \Pr_g(a \succ b \mid b \succ a) = 0.636$$

These probabilities can be used to compute the probability distribution over another ranking.

$$\Pr_o(a \succ b \mid a \succ b) = 0.347, \Pr_o(b \succ a \mid a \succ b) = 0.653$$

and

$$\Pr_o(b \succ a \mid b \succ a) = 0.664, \Pr_o(a \succ b \mid b \succ a) = 0.336$$

Finally we can compute the prediction-normalized votes using Equation (2).

$$\bar{V}(a \succ b) = 0.3 \cdot \left(1 + \frac{g(b \succ a \mid a \succ b)}{g(a \succ b \mid b \succ a)}\right) = 0.4 \cdot \left(1 + \frac{0.653}{0.336}\right) = 1.176$$

and

$$\bar{V}(b \succ a) = 0.6 \cdot \left(1 + \frac{g(a \succ b \mid b \succ a)}{g(b \succ a \mid a \succ b)}\right) = 0.7 \cdot \left(1 + \frac{0.336}{0.653}\right) = 0.909$$

Since $\bar{V}(a \succ b) > \bar{V}(b \succ a)$ the predicted ranking is $a \succ b$ which is also the correct ranking.

For the setting of two outcomes (i.e. two rankings $a \succ b$ and $b \succ a$), there is another way to determine the correct outcome. The correct outcome will be *surprisingly popular* i.e. it will be more frequent than predicted by the voters through the prediction question. Since $\Pr_s(a \succ b | a \succ b) = 0.4$ the frequency of $a \succ b$ is 0.4. On the other hand, the average prediction of $a \succ b$ is $0.4 \cdot \Pr_o(a \succ b | a \succ b) + 0.6 \cdot \Pr_o(a \succ b | b \succ a) = 0.341$. Hence ranking $a \succ b$ is *surprisingly popular* and the correct ranking.

The following result due to Prelec et al. [31], rephrased in our context, guarantees that the ground truth ranking will have the highest prediction-normalized vote count under the assumption that the highest posterior probability for ground truth ranking π will be assigned by a voter who observes noisy ranking π.

Theorem 1 ([31]). *Suppose the prior $\mathcal{P}$ and the signal distribution $\Pr_s$ are such that $\Pr_g(\pi|\pi) > \Pr_g(\pi|\pi')$ for all distinct rankings $\pi, \pi' \in \mathcal{L}(A)$. Then, we have that $\Pr[\pi^* \in \operatorname{argmax}_{\pi \in \mathcal{L}(A)} \bar{V}(\pi)] \to 1$ as $n \to \infty$.*

The assumption $\Pr_g(\pi|\pi) > \Pr_g(\pi|\pi')$ is known as self-predicting property and states that a voter with observed ranking π also believes that the true ranking with highest probability is π. This assumption is necessary for prediction-normalized votes to correctly identify the ground truth.

3 Elicitation Formats & Aggregation Rules

Note that the prediction requested from voter i, $\Pr_o(\cdot|\sigma_i)$, is a distribution over $m!$ rankings. Eliciting this prediction would undoubtedly place significant cognitive burden on the voter. Thus, our goal is to elicit partial vote and prediction information from the voters. Since eliciting numerical information is known to be difficult [6], we focus on eliciting ordinal information for prediction. We develop aggregation rules for recovering the ground truth from ordinal information and empirically evaluate the effectiveness of SP voting.

3.1 Elicitation Formats

We focus on two types of vote reports, and for each of them, two types of prediction reports. Below we provide formal explanations of these formats in the context of our model. In the next section, we provide example phrasings that were used to pose the various questions to the participants in our empirical study. Let r_i and q_i denote the vote and prediction reports submitted by voter i, respectively.

- *Top vote:* Voter i reports the top alternative in her observed noisy ranking, i.e., $r_i = \sigma_i(1)$.
  - *Top prediction:* Voter i estimates the most frequent alternative among the other votes, i.e. $q_i = \operatorname{argmax}_{a \in A} \sum_{\sigma \in \mathcal{L}(A): \sigma(1) = a} \Pr_o(\sigma|\sigma_i)$.
  - *Rank prediction:* Voter i estimates the ranking of the alternatives by their frequency among the other votes, i.e. the proportion of voters who rank them on top. In particular, the voter is asked to report $q_i \in \mathcal{L}(A)$ such that $\sum_{\sigma \in \mathcal{L}(A): \sigma(1) = q_i(x)} \Pr_o(\sigma|\sigma_i) \geq \sum_{\sigma \in \mathcal{L}(A): \sigma(1) = q_i(y)} \Pr_o(\sigma|\sigma_i)$ for all $x > y$.
- *Rank vote:* Voter i reports her entire observed noisy ranking, i.e., $r_i = \sigma_i$.

– *Top prediction:* Voter i estimates the alternative that appears most frequently in the top position of the other votes. Formally, this is equivalent to the top prediction in case of a top vote: $q_i = \text{argmax}_{a \in A} \sum_{\sigma \in \mathcal{L}(A):\sigma(1)=a} \Pr_o(\sigma|\sigma_i)$.

– *Rank prediction:* Voter i estimates the most frequent ranking among the other votes, i.e., $q_i \in \text{argmax}_{\sigma \in \mathcal{L}(A)} \Pr_o(\sigma|\sigma_i)$. Note that this is different from the rank prediction in case of a top vote.

These reports give rise to four elicitation formats, which we refer to as Top-Top, Top-Rank, Rank-Top, and Rank-Rank with the first component denoting the vote format and the second denoting the prediction format. As a benchmark, we use Top-None and Rank-None, where top and rank votes are elicited, respectively, but no prediction information is elicited.

ALGORITHM 1: SP Voting

Input: Information reports $\{r_i\}_{i\in[n]}$, prediction reports $\{q_i\}_{i\in[n]}$, and probabilities $\alpha > 0.5$ and $\beta < 0.5$.
for *each pair of alternatives* (a, b) **do**
  $\left(\{r_i^{(a,b)}, q_i^{(a,b)}\}_{i\in[n]}\right) \leftarrow$ Extract-Reports$(\{r_i, q_i\}_{i\in[n]}, (a, b), \alpha, \beta)$
  /* Signal 1 (resp. 0) corresponds to $a \succ b$ (resp. $b \succ a$). */
  $N_{a\succ b} = \left\{j : r_j^{(a,b)} = 1\right\}$
  $N_{b\succ a} = \left\{j : r_j^{(a,b)} = 0\right\}$
  $f(a \succ b) = \sum_i \mathbb{1}\left\{r_i^{(a,b)} = 1\right\} / (|N_{a\succ b}| + |N_{b\succ a}|)$
  $f(b \succ a) = 1 - f(a \succ b)$
  $g(a \succ b) = \frac{1}{|N_{a\succ b}|+|N_{b\succ a}|}\left(\sum_{i\in N_{a\succ b}} q_i^{(a,b)} + \sum_{i\in N_{b\succ a}} q_i^{(a,b)}\right)$.
$T \leftarrow \phi$ /* Create a tournament */
for *each pair of alternatives* (a, b) **do**
  /* Ties are broken u.a.r. */
  if $f(a \succ b) \geq g(a \succ b)$ **then**
    $T \leftarrow T \cup a \succ b$
  else
    $T \leftarrow T \cup b \succ a$
return T

3.2 Aggregation Rules

There are two difficulties in directly applying the SP algorithm of Prelec et al. [31] — maximizing $\tilde{V}(\pi)$ given in Equation (2) — in our setting. First, the effectiveness of the approach depends on how accurately functions f and g from Equation (2) match their expected values, which in turn depends on how large the number of voters is compared to the number of options among which the ground truth lurks. In our case, since the ground truth is one of $m!$ rankings, the approach would be ineffective unless each question is answered by a number of voters much larger than $m!$.[2]

In order to avoid the problem of $m!$ possible ground truth rankings, we instead determine the ground truth comparison of each of $\binom{m}{2}$ pairs of alternatives independently by applying the algorithm from Equation (2) on the relevant pairwise comparison data extracted from the reports of

[2]We consider a setting with 4 alternatives and there are 20 responses per question.

the voters. Algorithm 1 executes the SP algorithm on two possible comparisons between a and b – $a \succ b$ and $b \succ a$. As discussed earlier, we can compute the *prediction-normalized vote* for $a \succ b$ and $b \succ a$, denoted as $\bar{V}(a \succ b)$ and $\bar{V}(b \succ a)$ respectively, and choose whichever has higher normalized vote. However, as we are using surprisingly popular algorithm for two alternatives ($a \succ b$, and $b \succ a$) we can use a simpler version than explicitly computing the prediction normalized score [31]. In fact, the correct alternative is surprisingly popular i.e. actual frequency is more than the predicted frequency. In order to check if ranking $a \succ b$ is surprisingly popular we first compute $f(a \succ b)$, the true frequency of the order $a \succ b$ among the voters, which we approximate from the votes $\{r_i^{(a,b)}\}_{i\in[n]}$. We then compute the average prediction report $g(a \succ b)$ which is approximated from the prediction reports $\{q_i^{(a,b)}\}_{i\in[n]}$. Now,$a \succ b$ is the correct order between a and b if $f(a \succ b) \geq g(a \succ b)$. Otherwise, $b \succ a$ is the correct order between a and b.

The second challenge in applying the SP algorithm is that, even for comparing a pair of alternatives, Equation (2) requires cardinal prediction information whereas our input is ordinal. We propose a simple parametric model in which, for each elicitation format, we use two parameters, $\alpha \in (0.5, 1)$ and $\beta \in (0, 0.5)$, to convert ordinal pairwise predictions into cardinal pairwise predictions to be utilized by the SP algorithm. Algorithm 2 describes how to extract the relevant information about a pair (a, b) from the input. We first describe how to extract the relevant votes about the pair (a, b) from the input. Consider a vote r_i from voter i. If r_i is a rank over the alternatives, then we set $r_i^{(a,b)}$ either 1 or 0 based on whether $a \succ_{r_i} b$ or not. On the other hand, if r_i is just an alternative, then set $r_i^{(a,b)}$ to either 1 or 0 depending on whether the reported alternative equals a or b. If the top alternative r_i is neither a nor b, we just discard report of voter i for determining the order for the pair of alternatives a, and b.

In order to extract the relevant prediction report about the pair (a, b), note that q_i can be either a rank or an alternative, and we want to convert it to a probability estimate of $P(a \succ b | a \succ b)$ or $P(a \succ b | b \succ a)$. For ranked prediction report, the main idea is to check if $a \succ_{q_i} b$ or not, and then use this information to set a numerical value to the probability estimates. Thus, Algorithm 2 takes as input two additional parameters, α and β, which are used to determine the value of the probability estimates.[3]

If q_i is a rank over the alternatives, then we first choose either α or β depending on the value of $r_i^{(a,b)}$. Call this choice p. Then set $q_i^{(a,b)}$ either p or $1 - p$ based on whether $a \succ_{q_i} b$ or not. On the other hand, if q_i is just an alternative, then set $q_i^{(a,b)}$ to either p or $1 - p$ depending on whether the prediction equals a or b. Finally, in case the predicted alternative is neither a nor b, the prediction report doesn't make any distinction between the two alternatives, and we set $q_i^{(a,b)}$ to $1/2$.

Example continued. We now instantiate Algorithm 1 for the example presented Section 2. We also assume that the number of voters is large for ease of presentation. Recall that we assumed $a \succ b$ is the realized ranking. Then 0.4 fraction of voters receive signal $a \succ b$ and 0.6 fraction of voters receive signal $b \succ a$. We also consider a setting where the voters report *Rank vote* and *Rank prediction*.

Now consider a voter with signal $a \succ b$. The voter computes posterior probability over another voter's vote.

$$\Pr_o(a \succ b \mid a \succ b) = 0.347, \Pr_o(b \succ a \mid a \succ b) = 0.653$$

[3]We used the same values of α and β across all the agents. When q_i is a rank, one can use more parameters to set the numerical values of the probability estimates e.g. using the position of the alternatives a, and b in the rank q_i. However, we didn't find any improvement beyond using two parameters α and β.

The voter reports $q_i = b \succ a$ since $b \succ a$ receives higher probability in the posterior. Now consider a voter with signal $b \succ a$. The voter can compute posterior probability over another voter's vote.

$$\Pr_o(b \succ a \mid b \succ a) = 0.664, \Pr_o(a \succ b \mid b \succ a) = 0.336$$

Such a voter reports $q_i = b \succ a$ since $b \succ a$ receives higher probability according to the posterior belief.

We now execute Algorithm 1 with $\alpha = 0.65$ and $\beta = 0.4$. Consider a voter with vote $a \succ b$ and prediction report $b \succ a$. Algorithm 2 assigns $q_i^{(a,b)} = 1 - \alpha = 0.35$ (line 309). Similarly consider a voter with vote $b \succ a$ and prediction report $b \succ a$. Algorithm 2 assigns $q_i^{(a,b)} = \beta = 0.4$ (line 311). Therefore, the average prediction report for the ranking $a \succ b$ is $0.4 \cdot 0.35 + 0.6 \cdot 0.4 = 0.38$ which is less than 0.4 the frequency of vote $a \succ b$. Therefore, Algorithm 1 correctly determines the ranking $a \succ b$. This example also highlights why we restrict $a > 0.5$ and $\beta < 0.5$.

Note that applying our algorithm for comparing each pair of alternatives independently results in a tournament, which we use for two prediction tasks: predicting the top alternative in the ground truth ranking and predicting the entire ground truth ranking. For the former task, we select the alternative that defeats the maximum number of other alternatives in the resulting tournament, breaking ties uniformly at random, and consider the frequency of predicting the correct top alternative. For the latter task, we compute the Kendall-Tau distance of the tournament from the ground truth ranking.

Finally, note that there are no prediction reports for Top-None and Rank-None and we consider a natural extension of SP voting. In particular, for Top-None, SP voting returns an acyclic tournament comparing alternatives by their plurality scores, and for Rank-None, it returns the (potentially cyclic) majority preference tournament. We then select an alternative/ranking as described earlier.

4 Experiment Design

To test the effectiveness of SP voting for recovering the ground truth ranking with only ordinal elicitation, we conducted an empirical study by recruiting 720 participants (turkers) from Amazon Mechanical Turk (MTurk), a popular crowdsourcing marketplace. An average turker spent about 15 minutes to complete the survey. The survey was designed as follows.

Datasets. To generate questions with an underlying ground truth comparison of alternatives, we used three datasets from three distinct *domains*:

1. The *geography* dataset[4] contains 230 countries with their 2019 population estimates according to the United Nations.
2. The *movies* dataset[5] contains 15,743 movies with their lifetime box-office gross earnings.
3. The *paintings* dataset[6] contains 80 paintings with their latest auction prices.

[4]Retrieved from worldpopulationreview.com
[5]Retrived from boxofficemojo.com/chart/top_lifetime_gross
[6]Generously provided by the authors of Prelec et al. [31].

ALGORITHM 2: Extract-Reports

Input: Information reports $\{r_i\}_{i\in[n]}$, prediction reports $\{q_i\}_{i\in[n]}$, pair (a,b), and probabilities $\alpha > 0.5$, and $\beta < 0.5$.

for $i = 1, \ldots, n$ **do**

/* Extract information report */

if r_i *is a ranking* **then**

Set $r_i^{(a,b)} = \begin{cases} 1 & \text{if } a \succ_{r_i} b \\ 0 & \text{o.w.} \end{cases}$

else if r_i *is a top alternative and* $r_i \in \{a,b\}$ **then**

Set $r_i^{(a,b)} = \begin{cases} 1 & \text{if } r_i = a \\ 0 & \text{if } r_i = b \end{cases}$

else

Ignore (r_i, q_i) for determining $a \succ b$

/* Extract prediction report */

if q_i *is a rank* **then**

Set $q_i^{(a,b)} = \begin{cases} \alpha & \text{if } a \succ_{q_i} b \text{ and } r_i^{(a,b)} = 1 \\ 1-\alpha & \text{if } b \succ_{q_i} a \text{ and } r_i^{(a,b)} = 1 \\ 1-\beta & \text{if } a \succ_{q_i} b \text{ and } r_i^{(a,b)} = 0 \\ \beta & \text{o.w.} \end{cases}$

else if q_i *is a top alternative and* $q_i \in \{a,b\}$ **then**

Set $q_i^{(a,b)} = \begin{cases} \alpha & \text{if } q_i = a \text{ and } r_i^{(a,b)} = 1 \\ 1-\alpha & \text{if } q_i = b \text{ and } r_i^{(a,b)} = 1 \\ 1-\beta & \text{if } q_i = a \text{ and } r_i^{(a,b)} = 0 \\ \beta & \text{o.w.} \end{cases}$

else

Set $q_i^{(a,b)} = 1/2$.

return $\left(\{r_i^{(a,b)}, q_i^{(a,b)}\}_{i\in[n]}\right)$

We chose the three domains because they have varying levels of difficulty with *geography* being the easiest, and *paintings* being the hardest to answer. We want to see how various elicitation formats perform across questions of different levels of dificulty.

Questions. In each domain, the numerical values associated with the alternatives allow a ground truth comparison among the alternatives. For each domain, we considered the top 50 alternatives with the highest values. From these, we generated 20 questions, each comparing four alternatives selected such that two consecutive alternatives in the ground truth ranking were exactly 6 ranks apart in the global ranking of all 50 alternatives. Collectively, we had 60 questions across all three domains. For each of the 60 questions and each of the 6 elicitation formats described in Section 3, we elicited 20 responses, generating a total of 7,200 responses.

Turker Assignment. Figure 1 shows the workflow faced by a turker. Each of the 720 turkers responded to 10 questions split evenly among two randomly assigned elicitation formats. The turkers were divided roughly equally between the 30 ordered pairs of elicitation formats, which we call *treatments*. Further, as mentioned above, each question was assigned to the same number

of turkers under all elicitation formats.

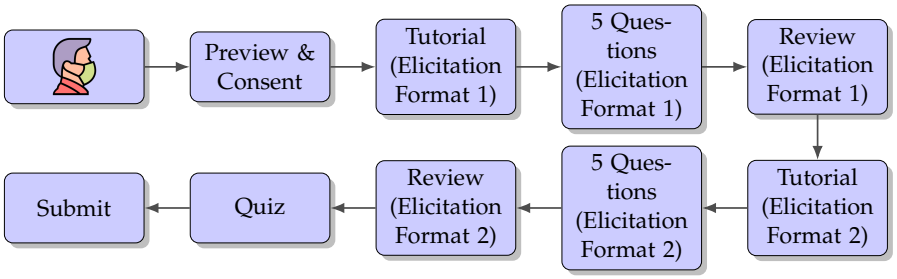

Figure 1: The workflow of a turker.

Tutorials. As shown in Figure 1, each set of five questions in a fixed elicitation format was preceded by a tutorial. The tutorial was designed specifically for the elicitation format and tested turkers' understanding of the vote and prediction formats. It contained a sample question along with pre-specified beliefs over the correct answer as well as over the other responses. Turkers had to successfully pass the tutorial by converting the given beliefs into the requested vote and prediction format in order to proceed to the questions.

Reviews. Each set of five questions was also succeeded by a review, which asked the turkers to rate the *difficulty* (from Very Easy to Very Difficult) and *expressiveness* (Very Little to Very Significant) of the elicitation format of the preceding questions. While we controlled the difficulty level of various questions from a given domain, as we show in Section 5, the three domains themselves differed significantly in their difficulty. In anticipation of this and to ensure that the turkers' implicit comparison between their two assigned elicitation formats is not influenced by the domains, the study was designed such that the sequence of domains encountered by a turker in the first five questions precisely matched that in the next five questions. See the appendix for details such as the consent form, the tutorial for each domain, the review, and other details.

Response Qualifications. To ensure high-quality responses, in addition to providing training in the form of tutorials, we restricted participation in our study to turkers who had (a) at least 90% approval rate on previous tasks, (b) at least 100 completed tasks, and (c) the region set to US East (us-east-1) on MTurk. [7] Additionally, at the end of the survey, the turkers were required to answer a quiz, which repeated the four alternatives from the last question they answered and asked them to identify the alternative they chose or ranked first in their vote. The turkers were incentivized to answer the quiz correctly (see below). In our case, over 82% of turkers passed the quiz.

Payments. The payment was divided into two parts. A *base* payment of 50¢ was provided conditioned on completing the entire survey including all tutorials, questions, and reviews. A *bonus* payment of 50¢ was provided conditioned on correctly answering the quiz question.

[7] The last restriction was set to ensure English proficiency and avoid language barriers.

Elicitation Formats. In Section 3, we discussed six elicitation formats and described what vote and prediction a given voter i should submit as a function of her observed noisy ranking σ_i, the prior $\mathcal{P}$, and the signal distribution $\Pr_s$. In our empirical study, we design natural and intuitive phrasing to elicit the corresponding responses from the turkers. As an example, consider a question which asks to compare four countries (United Kingdom, Vietnam, Russia, and Kenya) by their population.

1. Top-None: A turker is provided with four choices and is asked to vote the best option according to her opinion.

   - **Part A (vote):** *Which country do you think is the most populated among the following?*

   The turker is just asked one question and there is no additional prediction question regarding others' opinions.

2. Top-Top: A turker is asked two questions in this format. She is asked to vote her top choice as in format Top-None. Moreover, she is asked a prediction question about the votes of other participants.

   - **Part A (vote):** *Which country do you think is the most populated among the following?*
   - **Part B (prediction):** *Imagine that other participants will also answer Part A. Which of the following four countries do you think will be the most common response?*

3. Top-Rank: Like Top-Top, this rule also asks a turker two questions. However, the prediction question is different, and asks to rank the four choices based on the votes of other participants.

   - **Part A (vote):** *Which country do you think is the most populated among the following?*
   - **Part B (prediction):** *Imagine that other participants will also answer Part A. How do you think the following countries will be ordered from the most common response (top) to the least common (bottom)?*

4. Rank-None: This elicitation format asks the turker to order the four choices based on her own opinion.

   - **Part A (vote):** *How do you think the following four countries should be ordered from the most-populated (top) to the least-populated (bottom)?*

5. Rank-Top: This format also asks a turker to rank four choices and same as in Rank-None. Additionally, it asks the turker a prediction question about the votes of other participants.

   - **Part A (vote):** *How do you think the following four countries should be ordered from the most-populated (top) to the least-populated (bottom)?*
   - **Part B (prediction):** *Imagine that other participants will also answer Part A. In your opinion, which country will be the most common top choice?*

6. Rank-Rank: In this format, both the vote and prediction questions ask the turker to rank the four choices.

- **Part A (vote):** *How do you think the following four countries should be ordered from the most-populated (top) to the least-populated (bottom)?*
- **Part B (prediction):** *Imagine that other participants will also answer Part A. In your opinion, which will be the most common ordering of the following countries?*

Figures 2 and 3 show sample questions for different elicitation formats. Each turker completes five questions from two elicitation formats.

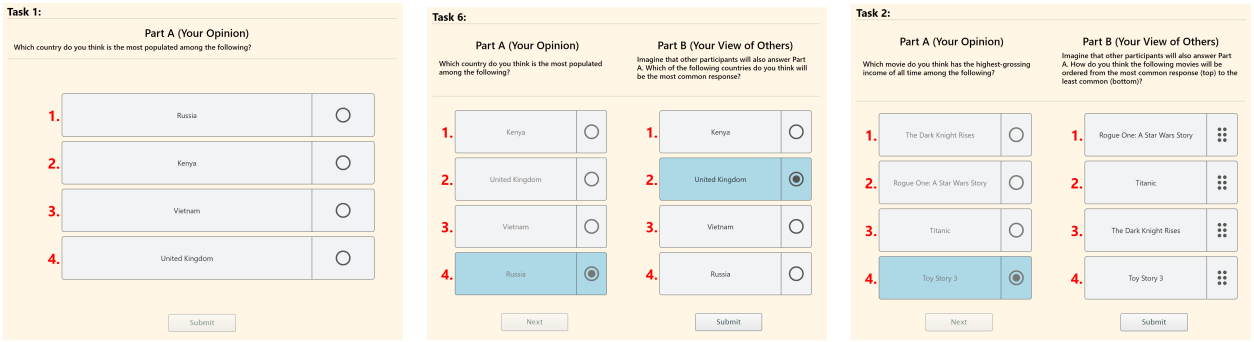

Figure 2: Sample questions from Top-None, Top-Top, and Top-Rank questions.

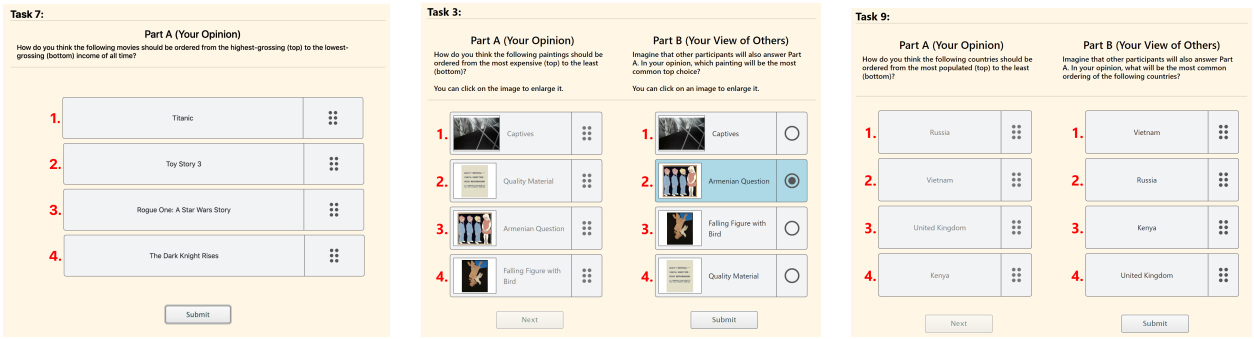

Figure 3: Sample questions from Rank-None, Rank-Top, and Rank-Rank questions.

Training. Recall that in our aggregation method, for each elicitation format, we use two parameters, $\alpha \in (0.5, 1)$ and $\beta \in (0, 0.5)$, to convert ordinal predictions into cardinal predictions that can be then used in SP voting (Algorithm 1). To learn effective values of these parameters, we split the dataset into a training and a test set. For each elicitation format, we selected 5 questions from each of three domains, reserving the remaining 15 questions from each domain for the test set. Using these 15 questions, we performed a grid search over α ranging from 0.55 to 0.95 in increments of 0.025 and β ranging from 0.05 to 0.45 in increments of 0.025 and selected the values with the lowest mean squared error.

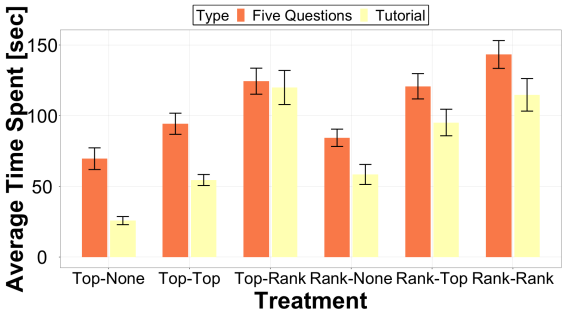

Figure 4: Average time spent on a single question.

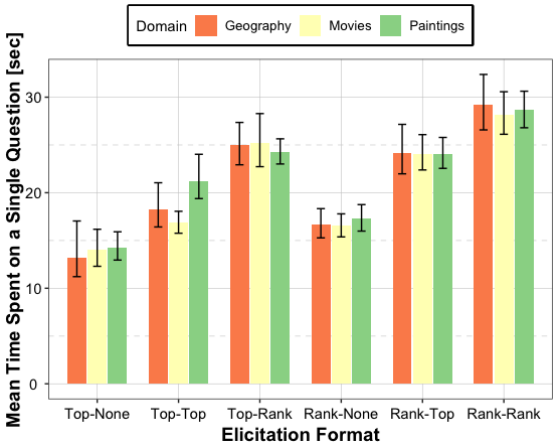

Figure 5: Average time-spent on a question, grouped by domain.

5 Results

In this section, we present our results averaged across all three domains. All confidence intervals shown are 95% intervals. We compare the elicitation formats using four key metrics: difficulty (i.e. cognitive burden), expressiveness, error in predicting the ground truth top alternative, and error in predicting the ground truth ranking.

5.1 Difficulty & Expressiveness

We measure the following three metrics.

- *Response time:* Response time is known to be a good objective proxy for the cognitive load associated with a task [32]. We measure the amount of time spent by the turkers on the tutorials and questions of the elicitation format.
- *Perceived difficulty:* As a subjective indicator of difficulty, we consider the perceived difficulty reported by the turkers (from Very Easy to Very Difficult) during the review stage of the elicitation format.
- *Perceived expressiveness:* Expressiveness indicates the amount of information that the turkers felt they were able to convey through the elicitation format (from Very Significant to Very Little).

Figure 4 shows the average time spent by the workers on the tutorial and on an average question under the six elicitation formats along with 95% confidence intervals (lower is better). We observe a statistically significant trend: when we fix a vote format (say Top or Rank), the average time spent increases for both tutorials and questions as we make the prediction format more complex (None $\rightarrow$ Top $\rightarrow$ Rank). Figure 5 shows the average time spent by a turker on a single question, for three different domains. For each domain, we observe the same pattern. For a fixed type of vote (either Top or Rank), as we ask more complex prediction reports (none $\rightarrow$ top $\rightarrow$ rank), the particular elicitation format requires more time to answer the questions.

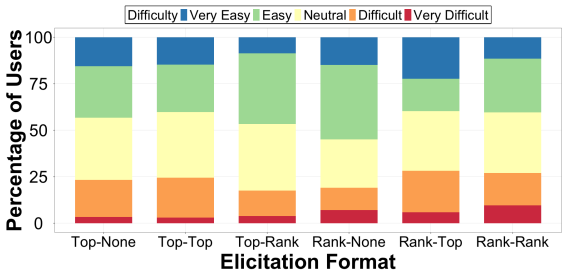

Figure 6: Perceived difficulty.

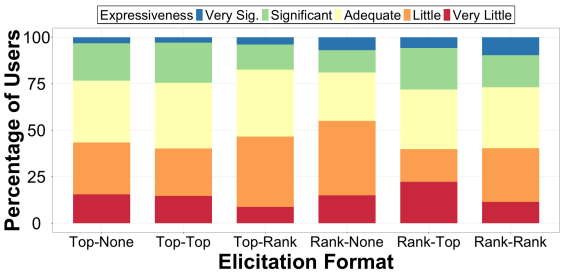

Figure 7: Perceived expressiveness.

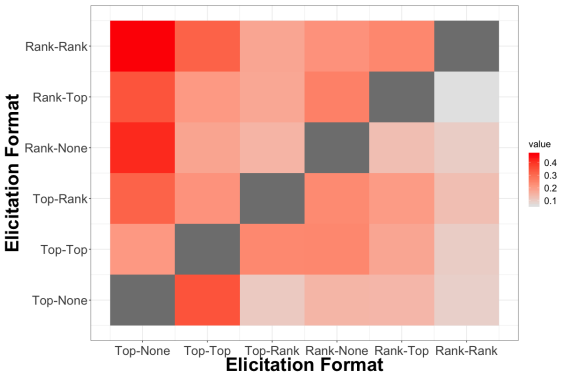

Figure 8: Relative difficulty.

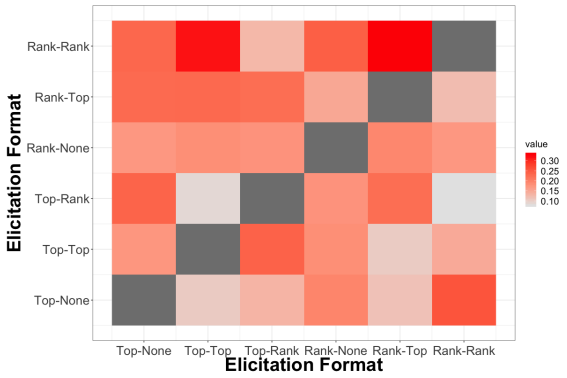

Figure 9: Relative expressiveness.

Figures 6 and 7 respectively show the reported distributions of perceived difficulty (easier is better) and perceived expressiveness (higher is better). Interestingly, the turkers found the six elicitation formats to be of very similar difficulty and similar expressiveness.

We also plot the relative difficulty (Figure 8) and relative expressiveness (Figure 9) among different elicitation formats. For a given treatment pair (t_1, t_2) the corresponding entry indicates how often a user reports elicitation format t_1 to be more difficult (or more expressive) compared to the elicitation format t_2. Note that, compared to Figures 6 and 7, the relative difficulty and expressiveness plot (Figures 8 and 9) provide an in-subject analysis and provide more information since we now observe a pairwise comparison plot rather than an aggregate plot for each elicitation format. In Figure 8 we see a clear trend – formats like Rank-Rank or Rank-Top are more difficult than formats with only Top votes e.g. Top-None or Top-Top. However, Figure 9 doesn't reveal a clear picture regarding the relative expressiveness of various elicitation formats.

5.2 Predicting the Ground Truth Top Alternative

We now turn to analyzing how effectively the different elicitation formats help us predict the ground truth. In addition to measuring the error of the ground truth estimate returned by our algorithm, we also measure the error in the input votes and predictions themselves. Note that every vote (and every prediction) is an estimate of some truth (either the ground truth or a summary statistic of the other votes); thus, its error can be measured with respect to the truth it is attempting to uncover.

First, we consider predicting simply the top alternative in the ground truth ranking. For our

algorithm as well as for the input votes and predictions, we use, as an error measure, the frequency of incorrectly guessing the top alternative of the truth they attempt to estimate. Figure 10 shows the average prediction errors for various elicitation formats (lower is better).[8] We remind the reader that the effectiveness of SP voting should be judged based only on elicitation formats which include some prediction information.

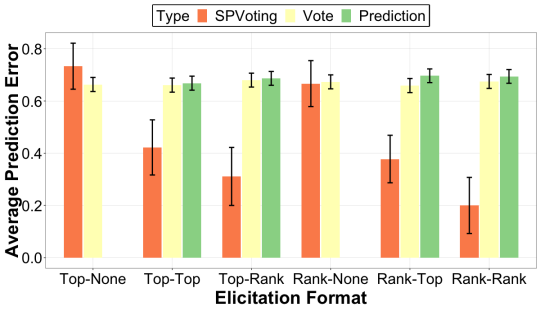

Figure 10: Average error in predicting the top alternative in the ground truth. By combining both the vote and predictions, SP voting achieves a much lower error than in either piece of information.

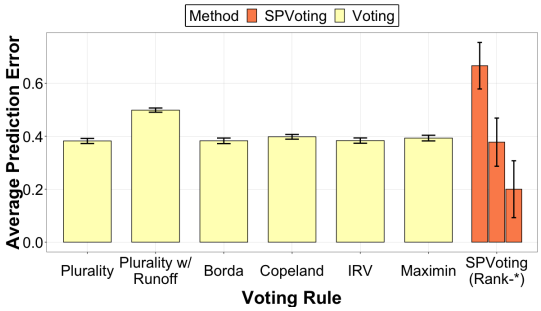

Figure 11: Comparing SP voting (Rank-None, Rank-Top, Rank-Rank respectively) with conventional voting for predicting the top alternative. By combining both vote and predictions, SP voting outperforms conventional voting.

Given four alternatives, selecting an alternative uniformly at random would result in a prediction error of 0.75. Interestingly, both the vote and prediction reports individually have average error around this benchmark. Yet, by combining these two pieces of individually erroneous information, SP voting is able to achieve significantly lower error. This is not surprising because SP voting approach is designed precisely to pick out the minority of experts lurking among a majority of non-experts by combining vote and prediction information. Moreover, for a fixed type of vote (either Top or Rank), as the prediction formats become more complex (None $\rightarrow$ Top $\rightarrow$ Rank), the performance of SP voting improves.

Figure 11 compares SP voting to several standard voting rules including Plurality, Plurality with Runoff, Borda, Copeland, Instant Runoff Voting (IRV), and Maximin Rule, which ignore the prediction information and simply aggregate the vote information in a democratic manner.[9] The conventional voting rules run on elections containing votes from three elicitation formats (Rank-None, Rank-Top, and Rank-Rank) whereas SP voting runs on each elicitation format individually. We can see that for Rank-Rank, SP voting (rightmost orange bar) outperforms all conventional voting rules, despite having access to just a third of the samples. This indicates that the prediction information helps significantly.

Figure 12 shows the average error in predicting the top alternative of the ground truth ranking for different elicitation formats and different domains. For each domain, we see that, for a fixed type of vote (Top or Rank) as we make the prediction reports more complex, the average prediction error generally goes down. In particular, except for the Paintings domain, the following orders always hold among the elicitation formats: Top-None $>$ Top-Top and Rank-None $>$ Rank-Top.

[8] SP voting errors are obtained by averaging over 60 elections associated with 60 questions. Vote/Prediction errors are averaged over 1200 responses and have narrower confidence intervals.

[9] See [5] for definitions of these rules.

Figure 13 compares our method with six conventional voting rules in terms of predicting the top alternative of the ground truth. We see the same phenomenon as we saw when all questions were combined. SP voting trained on just Rank-Rank elicitation format, outperforms all six voting rules for the domains Geography and Movies.

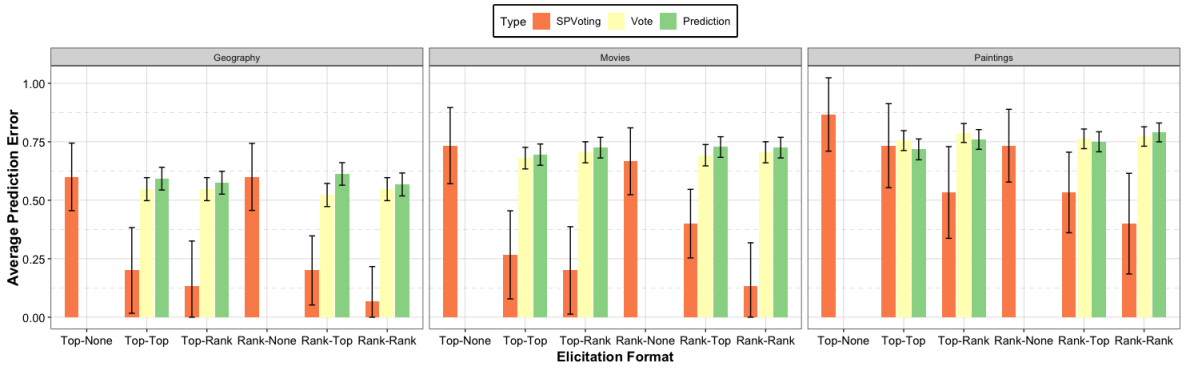

Figure 12: Average error in predicting the top alternative of the ground truth ranking, across different elicitation formats and different domains.

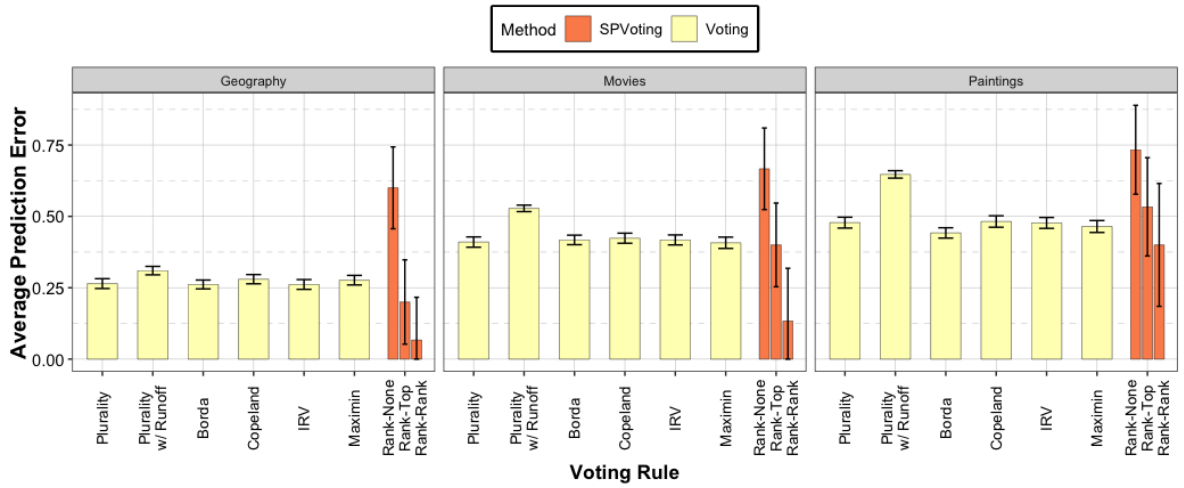

Figure 13: Average error in predicting the top alternative of the ground truth ranking, for different voting rules, and SP voting on three elicitation formats (Rank-None, Rank-Top, and Rank-Rank).

5.3 Predicting the Ground Truth Ranking

We now consider predicting the complete ground truth ranking. For SP voting result as well as the individual votes and predictions, we use the Kendall-Tau (KT) distance to measure the error of the SP voting result, votes, and predictions compared to the true ranking they aim to estimate. Figure 14 shows the average KT distance for different elicitation formats (lower is better). Given four alternatives, selecting a uniformly random ranking will have an average KT distance of 3. Both the votes and prediction reports have average error around this benchmark. Similar to predicting the top alternative, SP voting produces significantly lower average error by combining

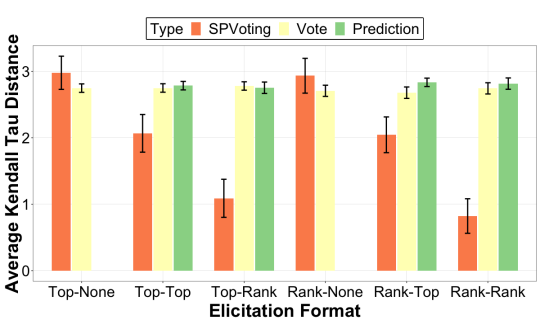

Figure 14: Average error in predicting the ground truth ranking. By combining both the vote and prediction information, SP voting achieves a much lower error than in either piece of information.

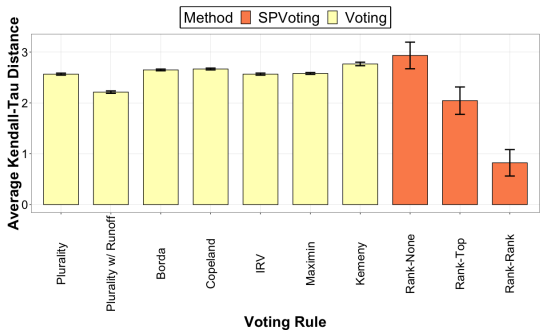

Figure 15: Comparing SP voting with conventional voting for predicting the ground truth ranking. Incorporating the prediction reports helps SP voting significantly outperform conventional voting.

these two noisy pieces of information. Moreover, for each vote format (either Top or Rank), as the prediction report becomes more expressive (None $\rightarrow$ Top $\rightarrow$ Rank) the average error of SP voting decreases.Finally, we also compare SP voting with standard voting rules (Figure 15) in terms of the average KT distance and find that SP voting again outperforms all voting rules for Rank-Rank. Figure 16 shows the average Kendall-Tau distance from the ground truth ranking for different elicitation formats and different domains. For each domain, we see that, for a fixed type of vote (Top or Rank) as we make the prediction reports more complex, the average prediction error generally goes down. In particular, except for the Paintings domain, the following orders always hold among the elicitation formats: Top-None $>$ Top-Rank and Rank-None $>$ Rank-Rank.

Figure 17 compares our method with six conventional voting rules in terms of the average Kendall-Tau distance from the underlying true ranking. We see the same phenomenon as we saw when all questions were combined. SP voting trained on just Rank-Rank elicitation format, outperforms all six voting rules for all the domains.

5.4 Prediction vs. Vote

Our results illustrate the importance of prediction in recovering the ground truth. While eliciting ranked votes and predictions (Rank-Rank) achieves the lowest error, an intriguing question arises when we seek to choose an elicitation format that provides a reasonable tradeoff between accuracy and difficulty/expressiveness. Figures 10 and 14 show that Top-Rank significantly outperforms Rank-Top while both formats are comparable in terms of response time, perceived difficulty, and perceived expressiveness. Thus, if we wish to choose an elicitation format slightly more complex than Top-Top, making the prediction more expressive is more promising than that of the vote. The same observation holds when comparing Top-Top versus Rank-None. This shows that when a tradeoff between more complex vote and more complex prediction is necessary, eliciting more complex prediction may be better.

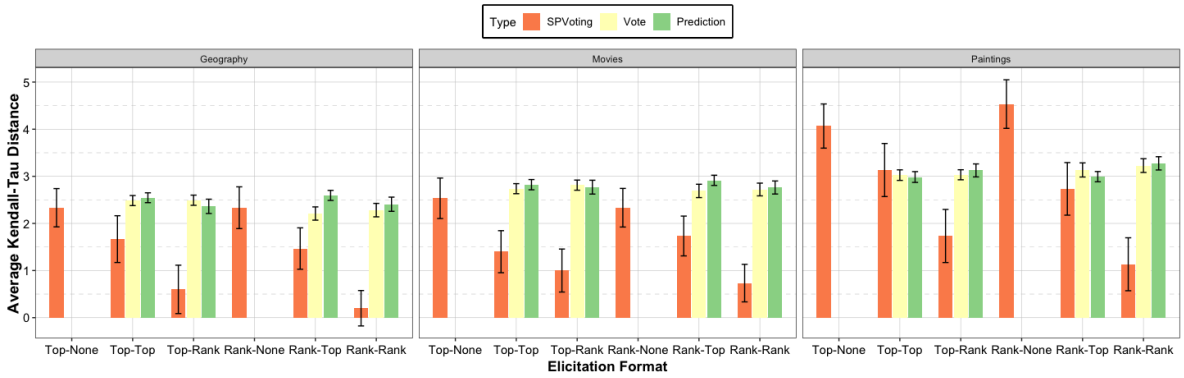

Figure 16: Average Kendall-Tau distance from the true rankings, across different elicitation formats, and different domains.

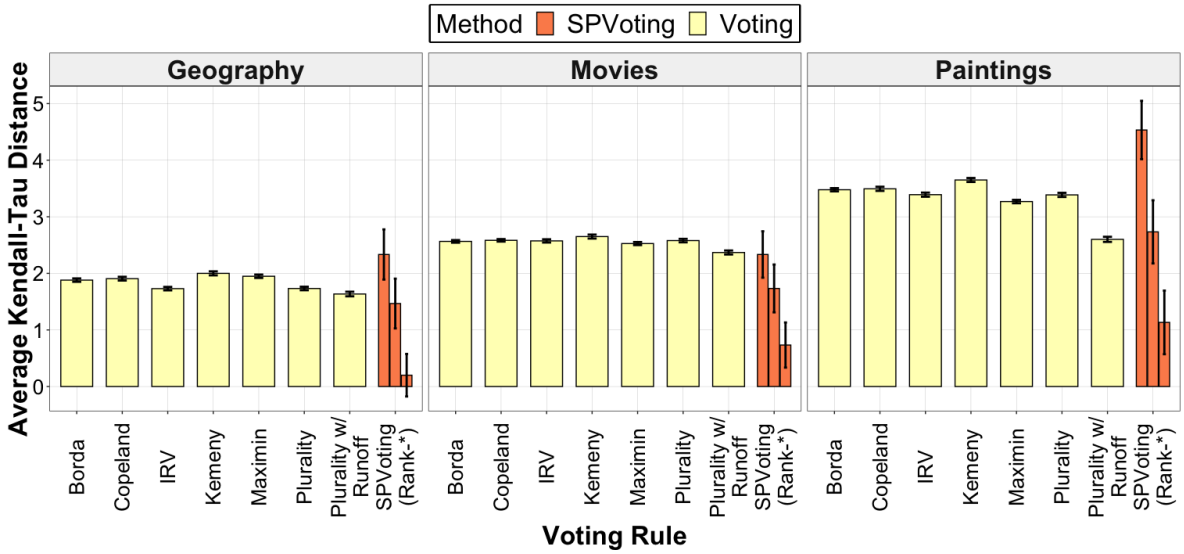

Figure 17: Average Kendall-Tau distance from the true rankings, for different voting rules, and SP voting on three elicitation formats (Rank-None, Rank-Top, and Rank-Rank).

6 Discussion

We extended surprisingly popular voting to recover a ground truth ranking of alternatives and, through a crowdsourcing study across different domains, showed that it outperforms conventional voting approaches without significantly increasing elicitation. In our study, the ground truth is a ranking over four alternatives, and a challenging future direction is to extend this approach to rankings with more than four alternatives. For a large number of alternatives, any practical elicitation scheme would ask the voters to report a partial rank over the alternatives, which will make it challenging to design aggregation rules for such partial ranks. A recent follow up work [17] has extended SPVoting to partial rankings by incorporating partial rank aggregation methods into the surprisingly popular framework.

Another interesting direction would be to derive theoretical performance guarantees for surprisingly popular voting when the number of participants is finite (the results of Prelec et al.

[31] hold only in the limit) and when only partial votes and predictions are elicited. In recent works [17, 18] we have derived sample complexity of SPVoting under the assumption of rank-order models (Mallows and Plackett-Luce) and clustering of expert and non-expert votes. However, it would be challenging to consider more general models (e.g. random utility models) and the effects of varying levels of expertise in the crowd.

Acknowledgements

The authors were partly supported by NSF grants #2144413 and #2107173 (Hosseini), a postdoctoral fellowship from Columbia DSI (Mandal), and an NSERC Discovery Grant (Shah).

References

[1] Ben Abramowitz, Elliot Anshelevich, and Wennan Zhu. Awareness of voter passion greatly improves the distortion of metric social choice. In *Proceedings of the 15th International Conference on Web and Internet Economics (WINE)*, pages 3–16, 2019.

[2] Arpit Agarwal, Debmalya Mandal, David C Parkes, and Nisarg Shah. Peer prediction with heterogeneous users. *ACM Transactions on Economics and Computation (TEAC)*, 8(1):1–34, 2020.

[3] Georgios Amanatidis, Georgios Birmpas, Aris Filos-Ratsikas, and Alexandros A. Voudouris. Peeking behind the ordinal curtain: Improving distortion via cardinal queries. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 1782–1789, 2020.

[4] Kenneth J. Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D. Nelson, George R. Neumann, Marco Ottaviani, Thomas C. Schelling, Robert J. Shiller, Vernon L. Smith, Erik Snowberg, Cass R. Sunstein, Paul C. Tetlock, Philip E. Tetlock, Hal R. Varian, Justin Wolfers, and Eric Zitzewitz. The promise of prediction markets. *Science*, 320(5878):877–878, 2008.

[5] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

[6] Colin F Camerer. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press, 2011.

[7] Ioannis Caragiannis, Ariel D Procaccia, and Nisarg Shah. When do noisy votes reveal the truth? *ACM Transactions on Economics and Computation (TEAC)*, 4(3):1–30, 2016.

[8] Yi-Chun Chen, Manuel Mueller-Frank, and Mallesh M Pai. The wisdom of the crowd and higher-order beliefs. *arXiv preprint arXiv:2102.02666*, 2021.

[9] Yiling Chen and David M Pennock. Designing markets for prediction. *AI Magazine*, 31(4): 42–52, 2010.

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

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

[12] Patrick M De Boer and Abraham Bernstein. Efficiently identifying a well-performing crowd process for a given problem. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1688–1699, 2017.

[13] Marquis de Condorcet. Essai sur l'application de l'analyse à la probabilité de décisions rendues à la pluralité de voix. Imprimerie Royal, 1785. Facsimile published in 1972 by Chelsea Publishing Company, New York.

[14] Boi Faltings and Goran Radanovic. Game theory for data science: Eliciting truthful information. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 11(2):1–151, 2017.

[15] Mirta Galesic, W Bruine de Bruin, Marion Dumas, A Kapteyn, JE Darling, and E Meijer. Asking about social circles improves election predictions. *Nature Human Behaviour*, 2(3):187–193, 2018.

[16] Francis Galton. Vox populi. *Nature*, 75:450–451, 1907.

[17] Hadi Hosseini, Debmalya Mandal, and Amrit Puhan. The surprising effectiveness of sp voting with partial preferences. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[18] Hadi Hosseini, Debmalya Mandal, and Amrit Puhan. Surprisingly popular voting with concentric rank-order models. In *Proceedings of the ACM on Web Conference 2025*, pages 3026–3036, 2025.

[19] David Kempe. Communication, distortion, and randomness in metric voting. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2087–2094, 2020.

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

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

[22] Steven P Lalley and E Glen Weyl. Quadratic voting: How mechanism design can radicalize democracy. In *AEA Papers and Proceedings*, volume 108, pages 33–37, 2018.

[23] Michael D Lee, Irina Danileiko, and Julie Vi. Testing the ability of the surprisingly popular method to predict nfl games. *Judgment and Decision Making*, 13(4):322, 2018.

[24] Debmalya Mandal, Ariel D Procaccia, Nisarg Shah, and David Woodruff. Efficient and thrifty voting by any means necessary. In *Advances in Neural Information Processing Systems*, pages 7180–7191, 2019.

[25] Debmalya Mandal, Goran Radanović, and David Parkes. The effectiveness of peer prediction in long-term forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2160–2167, 2020.

[26] Debmalya Mandal, Nisarg Shah, and David P Woodruff. Optimal communication-distortion tradeoff in voting. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 795–813, 2020.

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

[28] Asa B Palley and Jack B Soll. Extracting the wisdom of crowds when information is shared. *Management Science*, 65(5):2291–2309, 2019.

[29] Marcus Pivato. Realizing epistemic democracy. In *The Future of Economic Design*, pages 103–112. 2019.

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

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

[32] Matthias Rauterberg. A method of a quantitative measurement of cognitive complexity. *Human-computer interaction: Tasks and organisation*, pages 295–307, 1992.

[33] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of machine learning research*, 11 (4), 2010.

[34] Abraham M Rutchick, Bryan J Ross, Dustin P Calvillo, and Catherine C Mesick. Does the "surprisingly popular" method yield accurate crowdsourced predictions? *Cognitive research: principles and implications*, 5(1):1–10, 2020.

[35] Grant Schoenebeck and Biaoshuai Tao. Wisdom of the crowd voting: Truthful aggregation of voter information and preferences. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[37] Juntao Wang, Yang Liu, and Yiling Chen. Forecast aggregation via peer prediction. arXiv:1910.03779, 2019.

[38] H. P. Young. Condorcet's theory of voting. *The American Political Science Review*, 82(4):1231–1244, 1988.

[39] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *Advances in neural information processing systems*, 27, 2014.

A Screenshots from Our User Interface

In this section, we provide screenshots of different pages of our user interface.

A.1 Preview and Consent Form

Figure 18 shows the Preview and the Consent pages. After accepting our HIT, a turker first sees the Preview page. The turker needs to accept the consnt for participation before continuing with the tutorials.

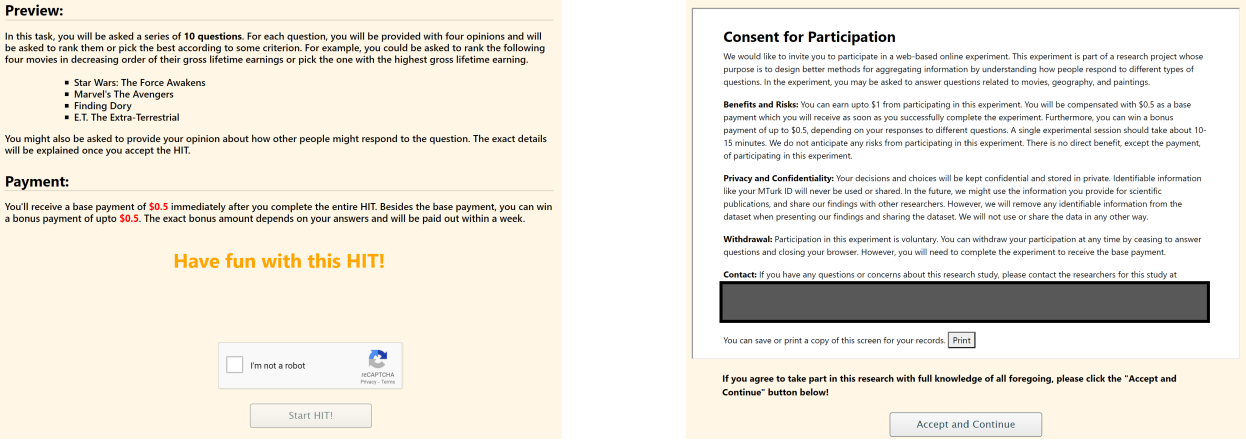

Figure 18: Preview and consent form

A.2 Tutorials

Figure 19 shows the tutorials for Rank-Top and Top-Rank elicitation formats. The tutorial provides the turker with a scenario (a correct answer, and a belief about other participants' votes), and asks the turker to complete the vote and prediction (if required) questions. Each turker must successfully complete the tutorial to proceed to the actual questions.

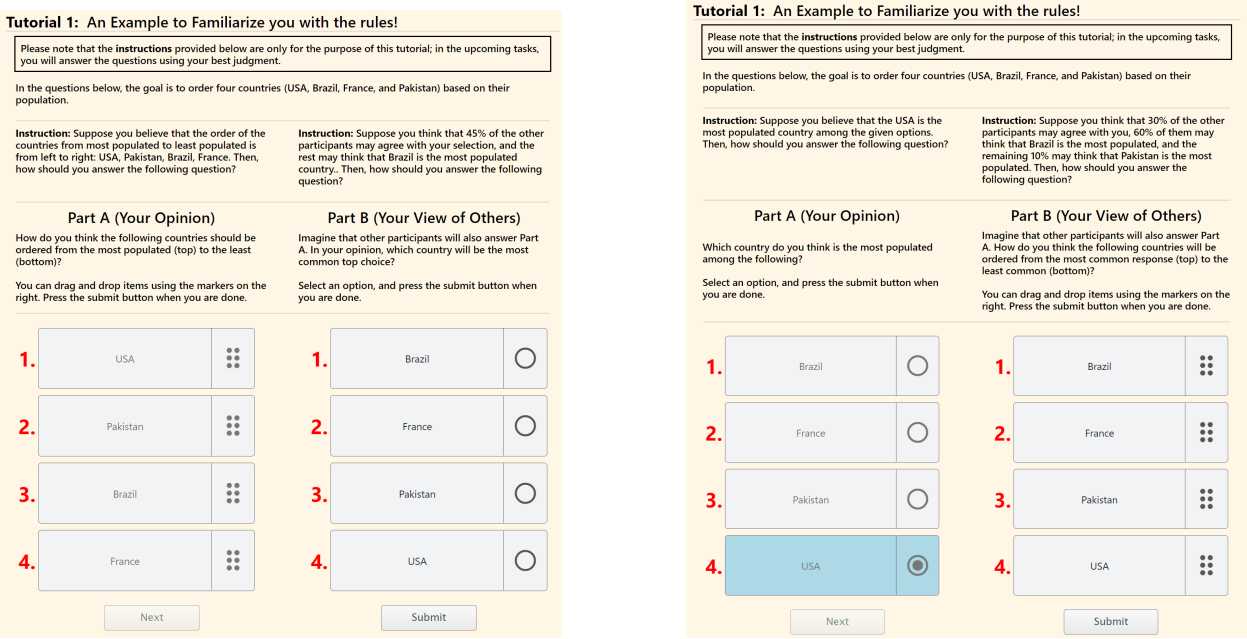

Figure 19: Tutorials for Rank-Top and Top-Rank formats.

A.3 Difficulty/Expressiveness

We assign each turker to two elicitation formats. After answering five questions from each format, the turker is asked to complete a survey about the difficulty and expressiveness of that format (shown in Figure 20).

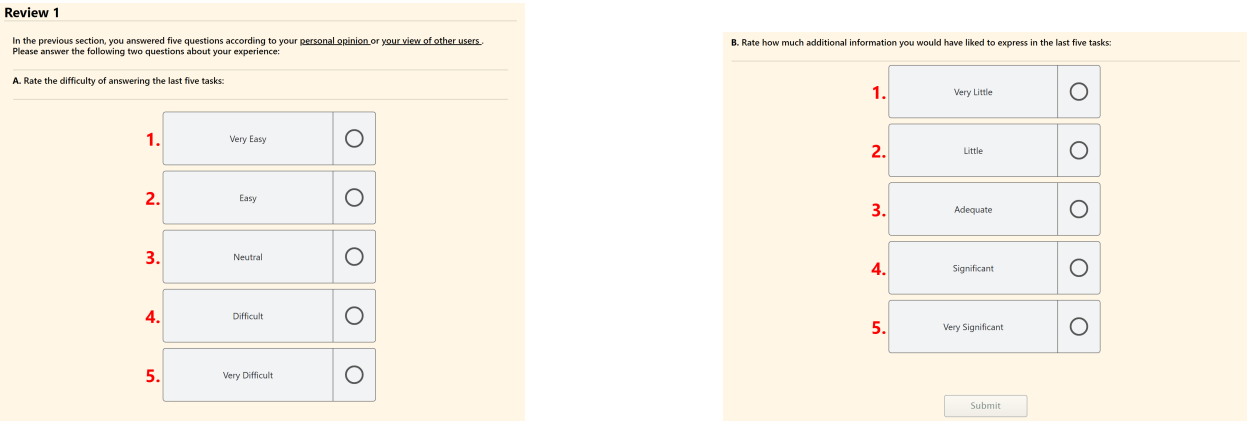

Figure 20: Difficulty and expressiveness questions