Incentives for self-extremized expert judgments to alleviate the shared-information problem*

Cem Peker[+]

[+]Erasmus University Rotterdam, peker@ese.eur.nl

May, 2021

Abstract

Decision makers frequently consult expert judgments. Simple average of forecasts from multiple experts is known to be effective in estimating uncertain quantities. However, benefits of averaging could be limited when experts have shared information, resulting in over-representation of the shared information in average forecast. This paper proposes a simple incentive-based solution to the shared-information problem. Experts are grouped with non-experts in forecasting crowds and they are rewarded for the accuracy of crowd average instead of their individual accuracy. In equilibrium, experts anticipate the over-representation of shared information and extremize their forecasts towards their private information to boost crowd accuracy. The self-extremization in individual expert forecasts alleviates the shared-information problem. Experimental evidence suggests that incentives for crowd accuracy induce self-extremization even in small crowds where winner-take-all contests (another incentive-based solution) are not effective.

*This research was made possible by European Research Council Starting Grant 638408 BayesianMarkets.

1 Introduction

Decision makers frequently require a reliable estimate/forecast of an uncertain quantity. Economists develop methods to nowcast or forecast economic indicators and make projections, which are essential for policy making (Elliott and Timmermann, 2013). Investors strive to predict future prices of commodities and assets accurately to make successful investments and achieve positive returns. Businesses invest vast resources into estimating demand for their existing and future products. Sports betting and election forecasting also involve predicting uncertain quantities (Stekler et al., 2010; Graefe et al., 2014).

Expert opinion could be a source of information to estimate uncertain quantities. Combining judgments of multiple experts typically produce accurate predictions (Armstrong, 2001). Aggregating multiple judgments incorporates decentralized and dispersed information held by a diverse group of individuals into a single estimate (Davis-Stober et al., 2014). The 'wisdom of crowds' effect occurs even for very small crowds (Mannes et al., 2014).

A decision maker who aims to utilize wisdom of crowds has to choose an aggregation method. Previous studies found simple averaging to be surprisingly effective and robust in a variety of estimation tasks (Genre et al., 2013; Clemen, 1989; Makridakis and Winkler, 1983; Mannes et al., 2012). When errors in individual judgments are statistically independent, simple averaging is effective in reducing errors in forecasting. Benefits of averaging could be limited when experts have shared information, which could result from an overlap in information sources (Gigone and Hastie, 1993; Chen et al., 2004). When best estimates of Bayesian experts are averaged, the shared information is over-represented in the aggregate estimate. As a result, the aggregate estimate exhibits the *shared-information bias* (Palley and Soll, 2019).

Recent work proposed aggregation mechanisms to address the shared-information problem. The pivoting method aims to recover the shared and private components of judgments and recombine them optimally (Palley and Soll, 2019). Knowledge-weighting proposes a weighted combination of judgments (Palley and Satopää, 2020). Pivoting and knowledge-weighting both rely on an augmented elicitation procedure where judges report their meta-predictions, i.e. a prediction on others' judgments (Prelec et al., 2017; Martinie et al., 2020; Wilkening et al., 2021). Pivoting requires meta-predictions to identify shared information while knowledge-weighting determines optimal weights based on the accuracy of meta-predictions. Another line of work suggests weighting judgments according to judges' expertise in similar estimation tasks to improve the aggregate estimate (Budescu and Chen, 2015; Mannes et al., 2014). Non-experts may rely more on shared information. Putting a lower weight on their judgments may reduce the undue influence of shared information in the crowd average. However, the shared-information bias persists even when non-experts are fully excluded because experts will also incorporate shared information in their predictions. Furthermore, such weighting methods are limited by the availability and reliability of past data.

This paper presents a simple incentive-based approach for aggregating judgments under shared information. We consider a setup where there is an unknown quantity and a sample of judges are asked to report a point estimate as a prediction. All judges observe a shared signal from the quantity while a subset of judges, referred to as experts, observe an additional private signal. Previous work on judgment elicitation typically uses proper scoring rules to elicit individuals' best estimates (Gneiting and Raftery, 2007). In contrast, we use a proper scoring rule to reward all individual predictions for the accuracy of the resulting crowd average. Under *incentives for crowd accuracy*, experts anticipate the shared-information problem and self-extremize towards their private signal to boost crowd accuracy. The self-extremization in individual expert judgments alleviates the shared-information bias in the average expert prediction. Unlike the alternative solutions discussed above, judges report a single point forecast only and no past data is required to determine weights for a weighted average of predictions.

We implement incentives for crowd accuracy in an experimental study very similar to Studies 1 and 2 in Palley and Soll (2019). Subjects are asked to predict the number of heads in 100 flips of a biased coin. All subjects observe a common sequence of sample flips, which represent the shared signal. Some subjects are assigned to 'expert' role. These expert subjects observe an additional judge-specific sequence of sample flips, which represent their private signal. Evidence suggests that experts self-extremize under incentives for crowd accuracy.

In presenting an incentive-based solution, we follow an approach similar to forecasting contests. In a winner-take-all contest of experts, an expert has an incentive to differentiate herself from others and avoid ties by adjusting her forecast towards her private information (Ottaviani and Sørensen, 2006; Lichtendahl Jr and Winkler, 2007; Pfeifer et al., 2014). As a result, the shared-information problem could become less severe (Lichtendahl Jr et al., 2013). However, the strength of incentives for self-extremization in a winner-take-all contest depends on the crowd size. In smaller crowds of experts, possibility of a tie (and hence, having to split the prize in the case of win) is lower. Then an expert would have weaker incentives to deviate from her best guess, making the contest less effective in correcting for the shared-information bias. We implement a winner-take-all contest of experts as an experimental condition in our studies. Results indicate that incentives for crowd accuracy achieve self-extremization in small crowds where winner-take-all incentives do not have a significant effect on predictions.

The rest of this paper is organized as follows: Section 2 introduces the formal framework and describes the shared-information problem. Section 3 develops incentives for self-extremization and establishes theoretical results. Section 4 presents experimental evidence. Section 5 provides a discussion of our findings and concludes.

2 The framework

2.1 Basics

The formal framework is similar to the specification of linear aggregation problem in Palley and Soll (2019). Let X be a random variable, which follows a known cumulative density $F(X|\theta)$ with unknown mean θ and a known finite variance. There are $N > 1$ risk-neutral Bayesian judges. Let $x \in \mathbb{R}$ be the ex-post realization of X. There is a decision maker who aims to elicit and aggregate the experts' judgments to estimate θ.

Judges share a common prior belief $\pi_0(\theta)$ on θ, where μ_0 and σ_0^2 are prior expectation and variance respectively. All judges observe the same common signal s_1, which is given by the average of m_1 independent observations of X. The sample of judges consist of $K \leq N$ experts $N - K$ laypeople, where $p = K/N$ represents the proportion of experts. Laypeople observe the common signal only. Experts both observe the common signal and receive a judge-specific private signal t_i, which is the average of ℓ independent observations of X. Without loss of generality, let judges $\{1, 2, \ldots, K\}$ be the experts. The special case $K = N$ corresponds to the symmetric information structure widely studied in the literature (Kim et al., 2001; Ottaviani and Sørensen, 2006; Lichtendahl Jr et al., 2013). The information structure and the parameters $\{K, N\}$ are common knowledge to the judges.

The information aggregation problem is *linear* if the posterior expectation of θ, given $F(X|\theta)$, is a linear combination of the prior expectation μ_0 and the signals $\{\mu_0, s_1, t_1, t_2, \ldots, t_K\}$ Palley and Soll (2019). In a linear aggregation problem,

$$E[\theta|\pi_0, s_1, t_1, t_2, \ldots, t_N] = \frac{m_0\mu_0 + m_1 s_1 + \ell \sum_{i=1}^{K} t_i}{m_0 + m_1 + \ell K}$$

where $E[\theta|\pi_0, s_1, t_1, t_2, \ldots, t_K]$ is referred to as the *global posterior expectation* (GPE). The GPE is the optimal aggregate forecast given the information provided by the common prior and the independent signals (Frongillo et al., 2015). Following Palley and Soll (2019), this paper considers X such that the information aggregation problem is linear [1]. In a linear aggregation problem, the prior mean μ_0 can be considered as representing m_0 observations of independent realizations of X. Let $m \equiv m_0 + m_1$ and $s \equiv (m_0 s + m_1 s_1)/m$. The *shared signal* s is a composite signal that represents the shared information of judges, consisting of the common prior and the common signal.

Using the simplified notation, the GPE can be written as follows:

$$E[\theta|s, t_1, t_2, \ldots, t_N] = \frac{m}{m + K\ell} s + \frac{\ell}{m + K\ell} \sum_{i=1}^{K} t_i \tag{1}$$

Each judge i updates her belief on θ after observing her signal (s, t_i) following Bayes' rule. It is common knowledge that judges are Bayesian. Let μ_i be the posterior expectation of judge i on θ. In a linear aggregation problem, we have

$$\mu_i = \begin{cases} (1-\omega)s + \omega t_i & \text{for } i \in \{1, 2, \ldots, K\} \\ s & \text{for } i \in \{K+1, K+2, \ldots, N\} \end{cases} \tag{2}$$

where $\omega = \ell/(m + \ell)$ is an expert's weight on the private signal. If judge i is a layperson, her posterior expectation is completely determined by the shared signal. An expert judge i's posterior expectation incorporates both the shared and private signals. The parameters (m, ℓ) are common knowledge to all judges.

2.2 The shared-information bias

Suppose each judge reports a point estimate x_i on X. Decision maker builds a crowd estimate by taking a simple average of individual reports. Let $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ be the crowd average. Consider the case where all judges report their true posterior expectations, i.e. $x_i = \mu_i$ for all $i \in \{1, 2, \ldots, N\}$. Let $\bar{x}_L = s$ and $\bar{x}_E = \frac{1}{K} \sum_{i=1}^{K} (1-\omega)s + \omega t_i$ denote the average

[1]See the online companion of Palley and Soll (2019) for examples of linear aggregation problems.

prediction of laypeople and experts respectively. Then, the crowd average can be written as

$$\bar{x} = (1-p)\bar{x}_L + p\bar{x}_E$$

Following Palley and Soll (2019), we define the *shared-information bias* as $E[\bar{x} - X|s, \theta]$, which can be written as follows:

$$\begin{aligned} E[\bar{x} - X|s,\theta] &= (1-p)E[\bar{x}_L - X|s,\theta] + pE[\bar{x}_E - X|s,\theta] \\ &= (1-p)(s-\theta) + p(1-\omega)((1-\omega)s + \omega\theta - \theta) \\ &= (1-p\omega)(s-\theta) \end{aligned} \tag{3}$$

The size of the shared-information bias depends on the proportion p of experts in the crowd, experts' weight ω on their private signal and the absolute difference between s and θ. Note that the bias exists even for $p = 1$. Each expert incorporates the shared signal in her prediction, resulting in an over-representation of shared information in average prediction even in crowds consisting of experts only. The bias does not disappear in large crowds for the same reason. The following section presents our solution to the shared-information problem.

3 Incentives for self-extremized expert judgments

In eliciting quantitative judgments, judges are typically rewarded for ex-post accuracy to motivate them to report their best estimates. Section 2.2 established that, when the judges report their best guesses on x, the crowd average exhibits the shared-information bias. This section develops *incentives for crowd accuracy*, where experts are rewarded for accuracy of the crowd average instead of their individual prediction. Then, expert's reports will not reflect their individual best estimates. Instead, we will show that experts put a higher relative weight on their private information. Such expert reports correct for the shared-information bias in the resulting average prediction.

The decision maker asks each judge i to report x_i simultaneously and aggregates estimates using $\bar{x}$. Let $C(\bar{x}, x)$ be the *crowd score* of the aggregate estimate $\bar{x}$, where C is a scoring function such that

$$x = \arg\max_{y \in \mathbb{R}} C(y, x) \tag{4}$$

$$\theta = \arg\max_{y \in \mathbb{R}} E[C(y, X)] \tag{5}$$

Intuitively, C is a measure of the ex-post accuracy of an estimate and the expected score is maximized at θ. All judges receive the same reward, determined according to $C(\bar{x}, x)$. Thus, the elicitation procedure motivates judges to report in a way that boosts the crowd accuracy. Let $\bar{x}_{-i}$ be the crowd average of all judges excluding i. The crowd average $\bar{x}$ can be written as follows:

$$\bar{x} = \frac{N-1}{N}\bar{x}_{-i} + \frac{1}{N}x_i$$

Then, judge i's expected payoff maximization problem can be expressed as follows:

$$\max_{x_i \in \mathbb{R}} E\left[C\left(\frac{N-1}{N}\bar{x}_{-i} + \frac{1}{N}x_i, X\right)\right] \tag{6}$$

Judges participate in a simultaneous reporting game where each judge i sets x_i to maximize the expected crowd score. Let x_i^* denote the optimal report of judge i.

How do incentives for crowd accuracy affect expert predictions? The theorem below presents an equilibrium of the simultaneous reporting game:

Theorem. *Under incentives for crowd accuracy, there exists a Bayesian Nash equilibrium in which*

$$x_i = \begin{cases} f^E(s, t_i) & \textit{for } i \in \{1, 2, \ldots, K\} \\ f^L(s) & \textit{for } i \in \{K+1, K+2, \ldots, N\} \end{cases}$$

where

$$\begin{aligned} f_E(s, t_i) &= \frac{m-(N-K)\ell}{m+K\ell} s + \frac{N\ell}{m+K\ell} t_i \\ f_L(s) &= s \end{aligned}$$

are the reporting strategies of experts and laypeople respectively. In this equilibrium experts self-extremize, i.e. they put a higher relative weight on the private signal.

Proof of the theorem is included in Appendix A. In equilibrium, laypeople simply report their posterior expectation on X. In contrast, experts' predictions differ from their posterior expectation given in equation 1. An expert weight on her private information in her best guess is $\omega = \ell/(m+\ell)$. Observe that the equilibrium weight on the private information, given by $N\ell/(m+K\ell)$, is strictly higher than ω. The opposite is true for experts' weight on the shared signal. Also note that $p = 1$ is simply a special case where all agents in the crowd put weights $m/(m+N\ell)$ and $N\ell/(m+N\ell)$ on the shared and private signals respectively. Under incentives for crowd accuracy, experts' anticipate the shared-information problem and self-extremize in their predictions by putting a higher relative weight on their private signal. The resulting $\bar{x}$ is given by

$$\begin{aligned} \bar{x} &= \frac{1}{N}\left\{\sum_{i=1}^{K} \frac{m-(N-K)\ell}{m+K\ell} s + \frac{N\ell}{m+K\ell} t_i + \sum_{i=K+1}^{N} s\right\} \\ &= \frac{m}{m+K\ell} s + \frac{\ell}{m+K\ell} \sum_{i=1}^{K} t_i \end{aligned}$$

In equilibrium, the crowd average reflects the GPE given in equation 1. Experts' self-extremizate such that the resulting crowd average does not exhibit the shared-information bias.

Would experts anticipate the shared-information bias and self-extremize in practice? Section 4 presents experimental evidence from two studies. In both studies, subjects are asked to predict the number of heads in 100 flips of a biased coin. Prior to making a prediction, subjects observe shared and private signals, which consist of independent sequences of sample flips. We implement incentives for crowd accuracy in treatment conditions to investigate if subjects' put a higher weight on their private signals, which would indicate self-extremization.

4 Experimental evidence

Section 3 established that when incentivized for crowd accuracy, Bayesian experts self-extremize towards their private information to correct for the shared-information bias. The result depends on experts' ability to anticipate the shared-information problem. In two experimental studies, we test if subjects are capable of such reasoning. Section 4.1 provides an overview of our experimental studies. Sections 4.2 and 4.3 provide a more detailed account of the designs, procedures and results.

4.1 Motivation and Overview

We run two controlled experiments to test if judges self-extremize under incentives for crowd accuracy. In both studies the experimental design is similar to studies 1 and 2 in Palley and Soll (2019). We recruit participants for an online experiment, in which subjects complete 10 prediction tasks. In each task, there is a two-sided coin with an unknown bias. Subjects are asked to predict the number of heads in 100 flips of the coin. Before making a prediction, subjects observe a shared signal consisting of 10 flips of the coin. In addition, some subjects receive an additional private signal which consists of another 10 flips from the same coin. After the experiment is completed we randomly pick one of the coins and flip it 100 times (virtually). Rewards are determined based on the outcome of these flips.

Study 1 is designed to make the shared information problem highly salient. Subjects are selected in forecasting crowds of sizes 5, 10 and 30. Each forecasting crowd of size N consists of one human subject and $N-1$ computer-generated (CG) agents. The CG agents predict based on the shared signal only. For example, if there are 7 heads out of 10 flips in the shared signal, all CG agents predict 70 heads in 100 flips. Each human subject is in the expert role (observes a private signal) and knows that the other crowd members are CG agents who predict based on the shared signal only. Each subject is rewarded according to the accuracy of her crowd's average forecast. The design also implements a control group where subjects in expert role are rewarded for their individual accuracy. We test if subjects self-extremize in the treatment conditions. Furthermore, we investigate if the crowd size has an impact on the self-extremizing behavior. In small crowds, subjects may not perceive the severity of the shared-information problem and self-extremize less. In larger crowds with many non-experts, the shared-information problem is more salient. However, an individual expert's report has a smaller effect on the crowd average, which may discourage self-extremization. The treatment conditions will show the extent of self-extremization in crowds of size 5, 10 and 30.

Study 2 implements a more realistic crowd accuracy condition where forecasting crowds are comprised of humans only. Subjects are assigned to expert and layperson roles specified in Section 2.1. Each expert is selected in a forecasting crowd where other members are laypeople peers. Unlike Study 1, experts do not have exact information on other crowd members' predictions. However, they could still anticipate that the other crowd members will heavily rely on the shared information. In addition, Study 2 includes a contest condition in which subjects in expert role participate in a winner-take-all contest. We compare the effectiveness of incentives for crowd accuracy and winner-take-all contests in inducing self-extremization.

4.2 Study 1 - self-extremization in expert predictions

Study 1 investigates self-extremization in a setup where the shared-information problem is easily recognizable for subjects. We also vary the crowd size to see if it has an impact on the effectiveness of incentives for crowd accuracy.

4.2.1 Design and Procedures

Task. Subjects are asked to predict the number of heads in 100 flips of a biased two-sided coin. There are multiple such coins and for each coin, probability of heads (the bias) is drawn uniformly from $[0.01, 0.99]$. The bias is unknown to subjects. Before submitting a prediction, subjects observe two sequences of 10 independent sample flips from the corresponding coin. The first sequence is common to all subjects and represents the shared signal. The second sequence is subject-specific and represents a subject's private signal. Then, subjects report a prediction by moving a slider on a scale 0 to 100. There are in total 40 such coins. Each subjects participates on 10 prediction tasks and hence, makes a prediction for 10 coins.

The prediction task represents linear aggregation problem with a binomial variable (Palley and Soll, 2019). The unknown bias in each coin corresponds to θ. Subjects predict the realization of X, which is a binomial random variable that represents the number of heads in 100 flips of the coin. Shared and private signals are 10 independent flips each, where each flip is a realization from a Bernoulli process. Since $m = \ell = 10$, the signals are equally informative and the Bayesian weight ω on the private signal in a judge's posterior expectation is 0.5.

Design. We construct a between-subjects design where two factors are manipulated to generate experimental conditions. The primary factor of interest is the incentivization scheme. In *individual accuracy* conditions, subjects are rewarded for the accuracy of their individual reports. In *crowd accuracy* conditions, we select each subject into a forecasting crowd where other members of the crowd are computer generated (CG) agents. In any given prediction task, the CG agents' predictions are completely determined by the shared signal. To illustrate, suppose the shared signal has 7 heads out of 10 flips. Then, all CG agents predict 70 heads in 100 new flips of this coin. Each forecasting crowd of size N includes $N-1$ such CG agents and 1 human subject. Subjects are informed about the composition of their crowd and the rule CG crowd members follow in their predictions. A subjects' payoff is determined by the average of all predictions (her report and $N-1$ CG predictions) in her crowd. We set three levels of crowd size, given by $N \in \{5, 10, 30\}$. Thus, there are in total four experimental conditions, which are denoted by {Individual, Crowd-5CG, Crowd-10CG, Crowd-30CG}. Figure 1 provides an example from the experimental interface in the Crowd-10CG condition.

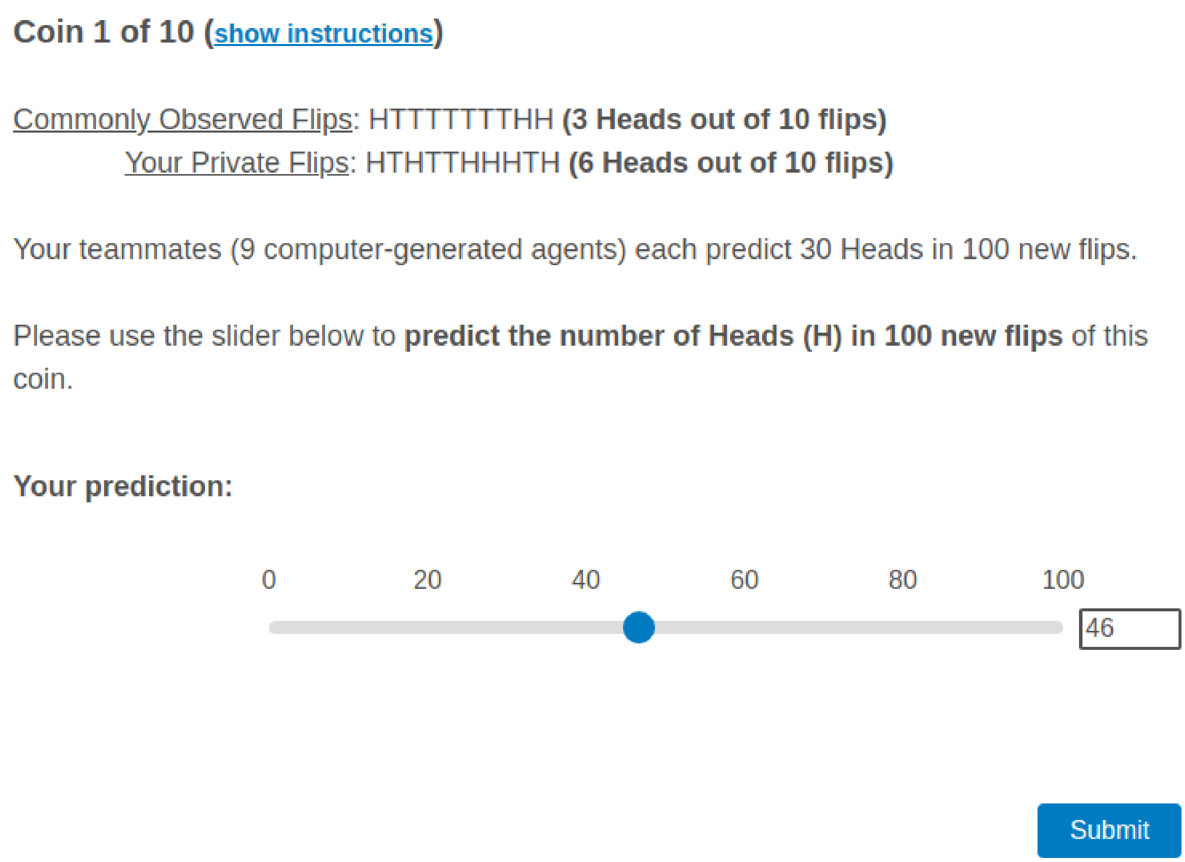

Figure 1: An example prediction task in the Crowd-10 condition. Initially, the slider starts at 0 and the text box that shows the current value is empty. The interface requires subjects to move (and release) the slider at least once or type a value directly.

As seen in Figure 1, subjects know that the predictions of other members of their crowd simply reflect the shared signal. This design makes the shared-information problem easily recognizable for subjects and allows us to test if subjects self-extremize in such a setting.

Subjects. Subjects are recruited from the online platform Prolific. We restrict the subject pool to students (at any level) who were US residents at the time of participation. The screening aims to recruit subjects who are more likely to understand the instructions and limit reporting errors. A total of 321 subjects completed the online experiment implemented via Qualtrics. Subjects are randomly assigned to one of the experimental conditions and spent on average 5 to 6 minutes to complete the experiment. Table B1 in Appendix B provides further information on the participants. For each coin used in the prediction tasks, we pre-generate the shared and private signals prior to the experiment. Each subject in a given condition observes a preset collection of shared and private signals. We use the same presets in each condition to improve the comparability of predictions across the experimental conditions.

Rewards. Subjects receive a participation fee of £1 for completing the experiment. In addition, they may earn a bonus based on their responses. After the experiment, we randomly pick a coin in each experimental condition and generate 100 flips. In the individual accuracy condition, subject i's bonus is calculated according to the bonus function B given as

$$B(x_i, x) = \begin{cases} 3 - \dfrac{3}{81}(x_i - x)^2 & \text{for } |x_i - x| \leq 9 \\ 0 & \text{for } |x_i - x| > 9 \end{cases} \tag{7}$$

where x_i is subject i's individual prediction and x is the realized number of heads in the 100 flips. The bonus function has a unique maximum at $x_i = x$. In the individual condition, B incentivizes subjects to report an estimate that minimizes the expected squared error, which corresponds to their posterior expectation on θ. Bonuses are positive for absolute forecasting errors smaller than 9. For example, if 38 heads appeared in 100 flips of the chosen coin and a subject predicted 33, her bonus is $3 - (3/81)5^2 =$ £2.07. The maximum bonus is £3 and bonuses never fall below 0.

Calculation of bonuses is similar in the crowd accuracy conditions, except that a subject's bonus is determined by accuracy of the crowd average. We calculate $\bar{x}^i$, which is the average of all predictions (subject i's prediction and $N - 1$ CG predictions) in subject i's crowd rounded to the closest integer. Then, subject i's bonus is determined according to $B(\bar{x}^i, x)$. Note that under incentives for crowd accuracy, B satisfies the conditions given in equations

4 and 5 for the scoring function C. The function $B(\bar{x}^i, x)$ has a unique maximum at $\bar{x}^i = x$ and the expected bonus $E[B(\bar{x}^i, x)]$ is maximized at $\bar{x}^i = \theta$ where the expected squared error is minimized. Subject i is incentivized to report x_i such that the resulting $\bar{x}^i$ reflects the GPE on θ, as in the theorem in Section 3.

Procedure. The online experiment is published on Prolific. Upon starting the experiment, subjects are selected into one of the experimental conditions. Then, subjects are presented with the instructions which explain the prediction task and rewards in the corresponding experimental condition. Explanation of the prediction task is identical across the conditions. Instructions are followed by a multiple choice quiz question about rewards. The quiz tests subjects' understanding of incentives for crowd or individual accuracy depending on the experimental condition and provides feedback to the subject before the tasks begin [2]. After the quiz, subjects are presented with the prediction tasks in a randomized order. Subjects complete the experiment by answering a few questions about their background and their experience in the experiment. Rewards are subsequently calculated and distributed on Prolific. Subjects' reports are retrieved from Qualtrics and matched with the data on demographics available through Prolific.

4.2.2 Results

We are interested in testing if incentives for crowd accuracy lead to self-extremization. Firstly, we follow Palley and Soll (2019) and compute *imputed weights*. For each subject i, we estimate the implied weight on private information from $(x_i - s)/(t_i - s)$. We exclude cases where the private signal is identical to the shared signal (i.e. same number of heads in sample flips). The imputed weights that do not fall in $[0, 1]$ are Winsorized: an imputed weight is set to 0 if it was negative and 1 if it was larger than 1. Recall that under the theorem in Section 3, subjects are predicted to extremize further in the direction of their private signals. We expect to see a higher frequency of cases where the imputed weight is

[2] All experimental data, instructions and quiz screens are available in the supplemental material.

1 under incentives for crowd accuracy. In contrast, we expect subjects in the individual condition to report their individual best guesses. Then the imputed weights would typically be 0.5. Figure 2 shows the distribution of imputed weights across all subjects and tasks in each experimental condition.

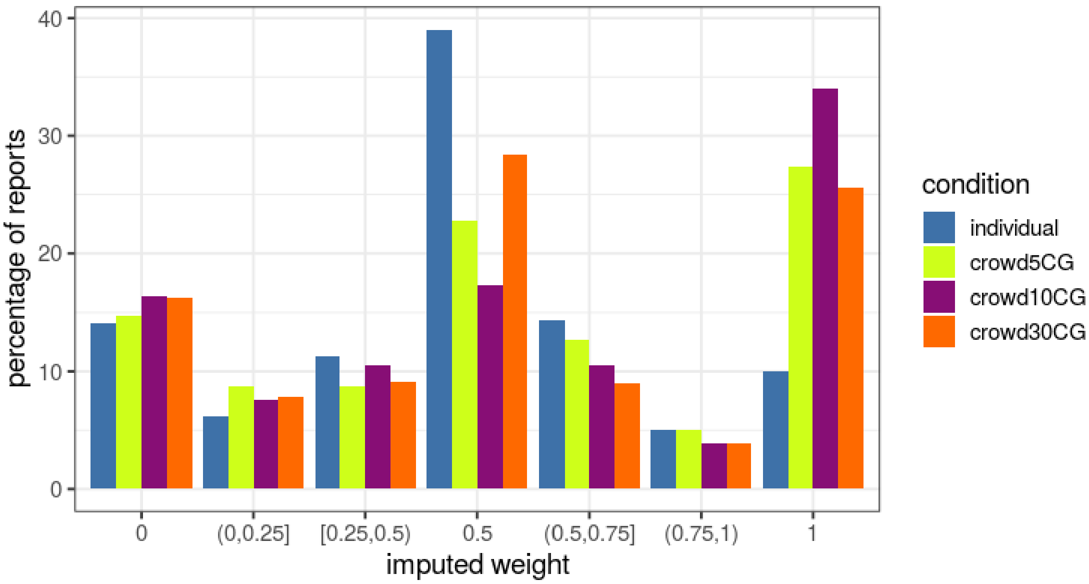

Figure 2: Frequency distribution of imputed weights

Most imputed weights in Figure 2 are 0, 0.5 and 1. Other values are binned in intervals of equal size. We see a clear difference between the individual accuracy and crowd accuracy conditions. As expected, the frequency of 0.5 is higher in the Individual treatment while the Crowd-5CG, Crowd-10CG and Crowd-30CG treatments show higher percentage of instances of 1. Figure 2 suggests that self-extremization occurs under incentives for crowd accuracy. Furthermore, self-extremization appears to be highest in the Crowd-10CG condition.

For further analyses, we consider *extremizing adjustment* as a measure of self-extremization. Consider a subject in the prediction task given in Figure 1. The shared signal suggests 30 heads in 100 new flips while the private signal suggests 60 heads. Since both signals are equally informative, a subject's posterior best guess is 45. Suppose subject i reported $x_i = 50$. We refer to $50 - 45 = 5$ as the extremizing adjustment. The subject's report falls 5 units closer to her private signal, which we take as an indication of self-extremization. In this example, the extremizing adjustment would be negative if the subject's report were less than 45. That would suggest a reporting behavior opposite of self-extremization.

Section 2.2 showed that the shared-information bias is proportional to $|s - \theta|$. Consider again Figure 1 and suppose $\theta = 0.5$ for that particular coin. The shared signal is simply 0.3 and we have $|s - \theta| = 0.2$. The shared-information bias becomes worse when s is further away from θ. We evaluate the extent of self-extremization in a prediction task by considering the *shared-information gap* given by $100 \times |s - \theta|$, where s and θ are the shared signal and the true probability of heads in that particular prediction task. Figure 3 depicts the average extremizing adjustment for various levels of the shared-information gap.

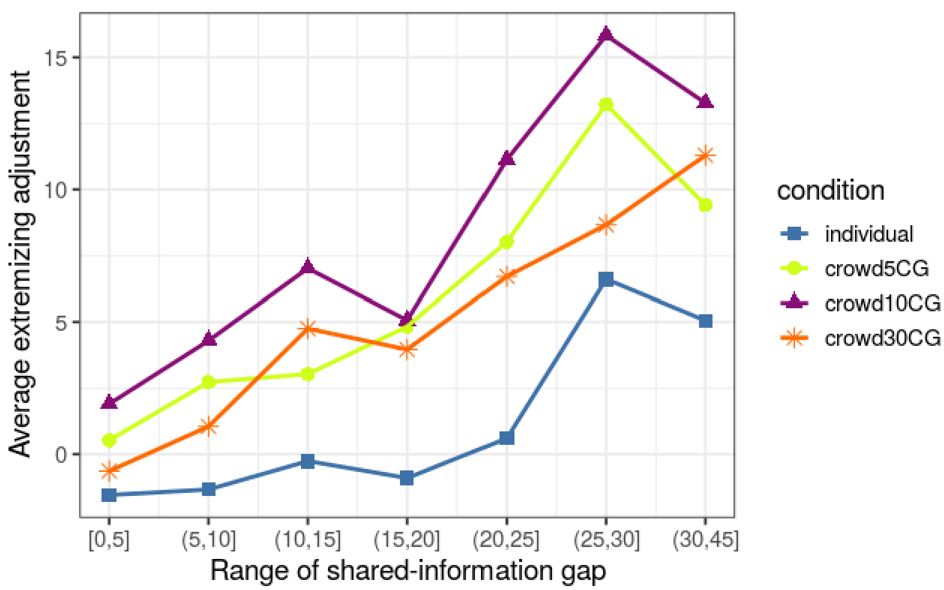

Figure 3: Average of extremizing adjustments for various levels of the shared-information gap

In the crowd accuracy conditions, average extremizing adjustment is positive and increasing in the shared-information gap. Recall that θ is also the expectation of the signal generation process. So, when the shared-information gap is higher private signals will on average substantially differ from s as well. Figure 3 suggests that subjects anticipate the benefits of self-extremization in particular when their private signal greatly differs from the shared signal. Interestingly, a positive level of extremizing adjustment occurs even in the

Individual condition when shared-information gap is sufficiently high. However, average extremizing adjustment is always higher in crowd accuracy conditions.

Table 1 below shows the estimates of the linear regression models where self-extremization rate is the dependent variable and the experimental condition is the independent variable of interest. The coefficients of Crowd-5CG, Crowd-10CG and Crowd-30CG measure the estimated difference in extremizing adjustments relative to the Individual condition. Model specifications (1) and (2) use the whole sample of subjects. In (3) and (4), subjects who gave an incorrect answer in the pre-experimental quiz or found instructions unclear are excluded to construct a filtered sample. Specifications (2) and (4) also include various controls. The variables 'US citizen?' and 'Female?' are binary indicators for US citizenship and gender respectively while 'Age' is a numeric variable. In all models, standard errors are clustered at subject level.

Table 1: Regression output. Standard errors are clustered at individual level.

Dep. var.: Extremizing adjustment				
	(whole sample)		*(filtered sample)*	
	(1)	(2)	(3)	(4)
(Intercept)	-0.28	2.69	-0.28	2.93
	(0.35)	(2.53)	(0.38)	(2.66)
Crowd-5CG	4.51^{***}	4.11^{***}	4.68^{***}	4.42^{***}
	(1.25)	(1.18)	(1.32)	(1.26)
Crowd-10CG	6.48^{***}	6.54^{***}	7.22^{***}	7.41^{***}
	(1.69)	(1.74)	(1.84)	(1.89)
Crowd-30CG	3.76^{***}	3.90^{***}	4.59^{***}	4.84^{***}
	(1.37)	(1.41)	(1.49)	(1.54)
Female?		-2.29^{*}		-2.33^{*}
		(1.23)		(1.32)
Age		-0.05		-0.05
		(0.10)		(0.10)
US citizen?		-0.39		-0.71
		(1.12)		(1.21)
R^2	0.02	0.03	0.03	0.03
Adj. R^2	0.02	0.02	0.02	0.03
Num. obs.	2601	2570	2362	2331
RMSE	16.41	16.42	16.19	16.19
N Clusters	321	317	292	288

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 1 shows significantly positive effects for all crowd accuracy conditions. Subjects extremize towards their private signal under incentives for crowd accuracy while the estimated extremizing adjustment is not different from zero in the individual accuracy condition (intercept term). Based on Table 1 and Figure 3 we can conclude that incentives for crowd accuracy induces self-extremization, which would alleviate the shared-information problem. Table 1 also suggests that the estimated extremizing adjustment is highest in Crowd-10CG. Pairwise tests of coefficients show no significant differences ($t = 0.97, p = 0.34$ in Crowd-10CG vs Crowd-5CG; $t = 1.28, p = 0.20$ in Crowd-10CG vs Crowd-30CG under model (1)). As discussed in Section 4.1, we might expect higher self-extremization in crowds of moderate size where experts would anticipate a serious shared-information problem while still being able to have a non-negligible effect on the crowd average through self-extremization. Despite the non-significance in differences between the crowd accuracy conditions, the higher estimated extremizing adjustment in the Crowd-10CG condition is in accordance with this intuition.

Results in Study 1 indicate that incentives for crowd accuracy elicit self-extremized expert predictions when the shared-information problem is highly salient. Study 2 further investigates incentives for crowd accuracy and provides a comparative analysis by implementing a winner-take-all contest as well.

4.3 Study 2 - Crowd accuracy vs winner-take-all contest

Study 2 uses the same prediction task as Study 1 but differs in two ways. Firstly, Study 2 implements incentives for crowd accuracy in a more realistic setting where all subjects including non-experts are humans. Secondly, Study 2 implements a winner-take-all contest of experts as another experimental condition. As discussed before, previous literature showed that subjects in a winner-take-all contest have incentives to self-extremize. We will compare incentives for crowd accuracy with winner-take-all contests in eliciting self-extremized expert predictions.

4.3.1 Design and Procedures

Task. The tasks in Study 2 are identical to those in Study 1. We use the same 40 coins and pre-generated shared and private signals to set up 40 prediction tasks. As in Study 1, each subject completes 10 prediction tasks.

Design. We follow a between-subjects design and manipulate incentivization scheme to generate three experimental conditions. The Individual condition is identical to the experimental condition of the same name in Study 1. We implement Individual in Study 2 as a benchmark. The experimental conditions of interest are Crowd-10 and Contest-10, which we explain below.

Recall that in Study 1, the estimated differences were highest in the Crowd-10CG condition. The Crowd-10 condition in Study 2 implements incentives for crowd accuracy in crowds of size 10. Unlike Study 1, forecasting crowds consists of human subjects only. Each subject is randomly assigned to the expert or layperson role, which they maintain in all tasks. An expert subject observes both the shared signal and a private signal while a layperson subject observes the shared signal only. Each forecasting crowd consists of 1 expert and 9 laypeople. The expert subjects are rewarded for the accuracy of their crowd average. In contrast, the layperson subjects are rewarded for their individual accuracy. Recall that in the equilibrium of the theorem, a layperson judge's prediction is simply the shared signal which is also her posterior expectation on the unknown quantity. So, the equilibrium report of a layperson judge under incentives for crowd accuracy is the same as her individual best estimate. In our experiment, rewarding layperson subjects for individual accuracy incentivizes them to predict according to the equilibrium reporting function f_L. Since we are only interested in experts' self-extremization, we reward layperson subjects for individual accuracy to keep the instructions simpler for both types of subjects. Experts are informed about the composition of their crowd. Unlike Study 1, experts do not know the exact predictions of the laypeople in the crowd. However, they know that the layperson subjects are incentivized to report their individual best estimates. Thus, experts could infer that laypeople will heavily rely on the shared signal.

In Contest-10 condition, each subject is in the expert role and participates in a winner-take-all contest with 9 subject. We split 40 prediction tasks in 4 'coin sets' of 10 tasks each. Experts in the Contest-10 condition complete one of the coin sets. Then, each expert in each set is selected into a group of 10 contestants, which consists exclusively of experts who completed the same set. After the experiment, we pick a coin randomly from each coin set and flip it 100 times to obtain the number of heads. An expert wins a bonus if her prediction on the chosen coin is the most accurate in her group of contestants. In case of a tie, bonus reward is split equally among the winners. We will provide more information on rewards below. The formation of coin sets and the assignment of experts to these sets are random. Similarly, experts are selected into contestant groups randomly. The tasks are organized in sets to ensure that subjects can be clustered in contestant groups of 10 for a randomly a chosen coin.

The Crowd-10 and Contest-10 conditions represent two incentive-based solutions to the decision maker's problem. Contest-10 is an implementation of a winner-take-all contest. An expert would like to report her best estimate to maximize her chances of winning the prize. However, the presence of shared information among contestants suggests the possibility of a tie and having to split the prize. Experts have an incentive to adjust their prediction away from the shared information. Crowd-10 groups experts with laypeople to elicit such self-extremized predictions. Experts anticipate the shared-information bias and self-extremize to improve the accuracy of crowd average, which determines their reward. The design allows a comparison on the effectiveness of incentives for crowd accuracy and a winner-take-all contest in eliciting self-extremized expert judgments.

Subjects. As in Study 1, we recruit subjects from Prolific and screen for students and US residents. In total, 295 subjects completed the experiment. Two subjects are excluded because their country of residence was different from the US. More information on subjects can be found in Table B2 included in Appendix B. In the Crowd-10 condition, the number of subjects that were assigned to the expert and layperson role are 81 and 47 respectively. The assignment of roles is set to be random until a sufficient number of layperson data is collected to construct crowds of 10 for each coin. As in Study 1, we are interested in experts' self-extremization. So, once we gathered sufficient layperson data, the incoming subjects are assigned to the expert role only.

Rewards. Participants receive £1 for completing the experiment. Bonuses in the Individual condition are calculated the same way as it is done in Study 1. Bonuses in the Crowd-10 condition are also similar to Study 1 and determined using the bonus function B

in equation 7. The layperson subject i's bonus is $B(x_i, x)$ where x_i is her prediction and x is the realized number of heads in 100 flips. An expert i's bonus depends on the accuracy of her crowd's average $\bar{x}^i$ and is given by $B(\bar{x}^i, x)$. In the Contest-10 condition, we calculate the absolute prediction error for each subject. For example, if $x = 60$ and subject i predicted 58, her absolute error is 2. A subject wins a bonus of £18 if she has the lowest absolute error in her contestant group. The prize is split evenly if 2 or more subjects are tied in being winners. Subjects who do not achieve the lowest absolute error in their group do not receive a bonus. The winner's prize is determined such that the expected bonus for an optimally self-extremizing expert (according to the theorem) in the Crowd-10 condition is equivalent to the expected bonus of a contestant in the Contest-10 condition. The resulting average bonuses for an expert in the Crowd-10 and Contest-10 conditions are £1.27 and £1.78 respectively. The ex-post discrepancy suggests that experts might have insufficiently self-extremizated for the corresponding levels of the shared-information bias in a crowd with 9 laypeople and 1 expert only. Note that the total prize in a contest is fixed, so the average bonus in Contest-10 does not depend on experts' self-extremization.

Procedure. Similar to Study 1, the online experiment is made available on Prolific. Incoming subjects are randomly selected into one of the three experimental conditions. Since the analysis is focused on expert judgment, the data collection is aimed at collecting approximately equal number of expert data across the experimental conditions. Recall that in the Crowd-10 condition, subjects are assigned to expert and layperson roles. In order to obtain more expert judgments in Crowd-10, we continued collecting expert data for Crowd-10 condition after Individual and Contest-10 conditions are stopped. Similar to Study 1, subjects see the instructions and complete a quiz. Explanation of the tasks is the same for the Individual and Contest-10 conditions as well as the expert role in Crowd-10. Layperson subjects in Crowd-10 observe the shared signal only. Thus, the instructions and the task interface do not include private signals. After the quiz, subjects complete prediction tasks in a randomized order and finish the experiment by completing a short survey (same as Study

1) on their background information and clarity of instructions. Rewards are calculated and distributed on Prolific.

4.3.2 Results

Similar to Study 1, we analyze extremizing adjustments in each experimental condition. Since layperson subjects in the Crowd-10 condition observe the shared signal only, self-extremization is not relevant. Figure B1 in Appendix B suggests that layperson subjects' predictions typically reflect the shared signal as predicted by the theorem. Figure 4 below shows the average extremizing adjustments of experts in each experimental condition for various levels of the shared-information gap.

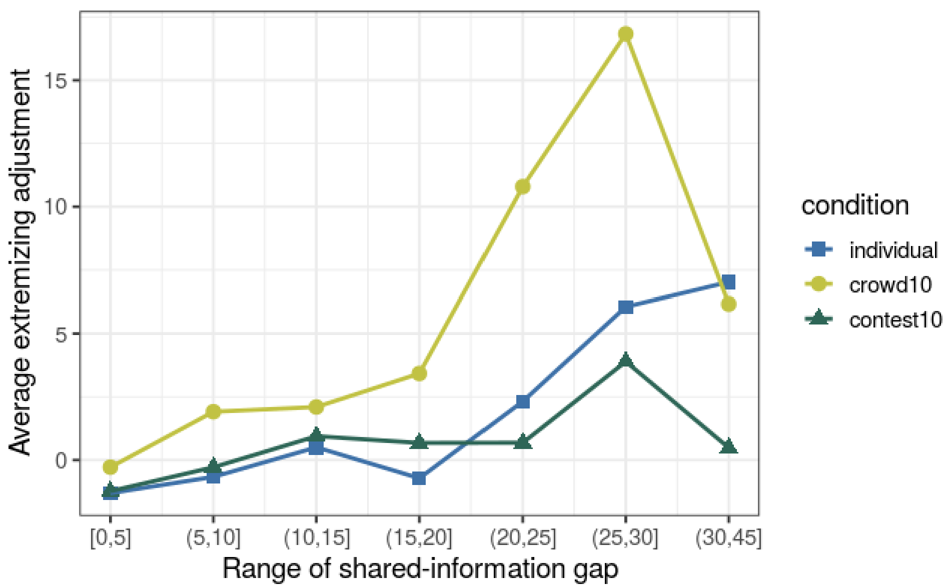

Figure 4: Average of extremizing adjustments for various levels of the shared-information gap

The average extremizing adjustment is always higher in Crowd-10 compared to Contest-10. In the Individual and Crowd-10 conditions, the extent of self-extremization is similar to the findings in Study 1. However, Figure 4 suggests that a winner-take-all contest with 10 experts does not induce significant levels of self-extremization. Table 2 presents the regression output. The coefficients of Crowd-10 and Contest-10 show the estimated extremizing adjustment above the level observed in the Individual condition. As in Table 1, models (1) and (2) use the whole sample while (3) and (4) filters the sample based on the quiz responses and self-reported understanding of the experiment. The controls are the same as before.

Table 2: Regression output. Standard errors are clustered at individual level.

Dep. var.: Extremizing adjustment				
	(whole sample)		*(filtered sample)*	
	(1)	(2)	(3)	(4)
(Intercept)	0.44	4.90***	0.29	5.19***
	(0.53)	(1.59)	(0.46)	(1.85)
Crowd-10	3.36**	3.44**	3.96***	4.16***
	(1.41)	(1.46)	(1.50)	(1.57)
Contest-10	−0.21	−0.49	−0.10	−0.22
	(0.68)	(0.65)	(0.69)	(0.70)
Female?		−0.97		−1.01
		(0.93)		(1.07)
Age		−0.09*		−0.12**
		(0.05)		(0.05)
US citizen?		−1.94		−2.03
		(1.18)		(1.41)
R^2	0.02	0.02	0.02	0.03
Adj. R^2	0.01	0.02	0.02	0.03
Num. obs.	1996	1978	1668	1668
RMSE	13.09	13.00	13.21	13.17
N Clusters	246	244	206	206

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 1 suggests a significantly higher level of extremizing adjustment in Crowd-10. In contrast, there are no differences between the Contest-10 and Individual conditions in terms of self-extremization. A pairwise comparison of Crowd-10 and Contest-10 also confirms that Crowd-10 elicits higher self-extremization ($t = 2.61, p = 0.009$ in Crowd-10 vs Contest-10 under model (1)).

To illustrate the implications of the results in Table 1, consider a decision maker who elicits predictions from a panel of 10 experts and considers a simple average of predictions as the aggregate prediction. In order to address the shared-information problem, the decision maker may consider setting up a winner-take-all contest to obtain self-extremized individual predictions. Table 1 indicates that such a contest does not elicit self-extremization beyond the levels observed when each expert is rewarded for accuracy individually. Instead, consider a setup where the decision maker has access to non-expert predictions, which are more heavily influenced by the shared information. Then, the decision-maker can elicit self-extremized expert judgments if she teams up each expert with a group of non-experts and rewards the expert based on the accuracy of her crowd's average.

5 Discussion

In extracting the wisdom of crowds, simple averaging of expert judgments has an intuitive appeal. The decision maker need not worry about identifying better experts, which is not a trivial task. Furthermore, evidence shows that simple averaging is hard to beat in many applications, implying a robustness across various information structures and application domains. However, simple average exhibits the shared-information bias when experts have shared information (Palley and Soll, 2019). In such cases, a decision maker would prefer experts to extremize their judgments away from the shared information. We proposed incentivizing experts for crowd accuracy as a means to elicit such judgments. The theory predicts that Bayesian experts would anticipate the shared-information problem and self-extremize to improve the accuracy of the crowd average. In two experimental studies we investigated if such self-extremization occurs in practice.

Study 1 essentially tests if experts follow the best response in the theorem given layperson predictions. The results in Table 1 imply that experts respond to incentives for crowd accuracy. Evidence from Study 2 suggests that expert self-extremization can emerge as an equilibrium when an expert is included in a crowd of non-experts. Even when experts do not know that other crowd members' predictions simply reflect the shared signal, they can anticipate that non-experts will heavily rely on the shared information and self-extremize in their prediction.

The practical effectiveness of incentives for crowd accuracy in eliciting self-extremized expert judgments depends on the salience of the shared-information problem. Unlike other solutions that elicit experts' meta-beliefs (Palley and Soll, 2019; Palley and Satopää, 2020) or use past data (Budescu and Chen, 2015; Mannes et al., 2014), we rely on Bayesian experts' ability to anticipate the shared-information problem. Incentives for crowd accuracy simply aligns the incentives of experts such that self-extremization increases their expected payoff. Previous work found mixed results in whether people have the correct intuition on the shared information and the resulting correlation between judgments (Soll, 1999; Budescu and Yu, 2007; Yaniv et al., 2009). In Study 2, we constructed forecasting crowds where the experts knew that their peers are non-experts. The results suggest that the experts understood why they should self-extremize. In a crowd of experts only or in more mixed crowds, the shared-information problem will be more subtle and the experts may not have the correct intuition. However, as in Study 2 the decision maker can select each expert into a crowd with non-experts and implement incentives for crowd accuracy in prediction problems experts and non-experts can be distinguished. Then the decision maker can elicit a self-extremized prediction from each expert. As shown in Section 3, the shared-information bias persists even in the average prediction of experts only. Averaging self-extremized expert predictions will alleviate the shared-information problem in the average expert prediction.

Study 2 also implemented a winner-take-all contest as an alternative incentive-based solution to elicit self-extremized expert judgments. Lichtendahl Jr et al. (2013) derived the pure strategy limiting equilibrium in a winner-take-all contest where experts self-extremize. The resulting average forecast is more accurate than the average of non-extremized forecasts. Pfeifer et al. (2014) illustrates why predicting the expert behavior in a finite sample of experts is challenging. The pure strategy equilibrium of self-extremization may not exist. Intuitively, motivation to self-extremize stems from experts' trade-off between reporting her best prediction and standing out from the others to avoid ties. In small samples, an expert's incentive to differentiate her forecast is weaker as a tie is much less likely. In our winner-take-all contests of 10 experts, we did not find any indication of self-extremization. Incentives for crowd accuracy present a solution to elicit self-extremized judgments in a small sample of experts.

References

Armstrong, J. S. (2001). Combining forecasts. In *Principles of forecasting*, pages 417–439. Springer.

Budescu, D. V. and Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2):267–280.

Budescu, D. V. and Yu, H.-T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, 20(2):153–177.

Chen, K.-Y., Fine, L. R., and Huberman, B. A. (2004). Eliminating public knowledge biases in information-aggregation mechanisms. *Management Science*, 50(7):983–994.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4):559–583.

Davis-Stober, C. P., Budescu, D. V., Dana, J., and Broomell, S. B. (2014). When is a crowd wise? *Decision*, 1(2):79.

Elliott, G. and Timmermann, A. (2013). *Handbook of economic forecasting*. Elsevier.

Frongillo, R. M., Chen, Y., and Kash, I. A. (2015). Elicitation for aggregation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Genre, V., Kenny, G., Meyler, A., and Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1):108–121.

Gigone, D. and Hastie, R. (1993). The common knowledge effect: Information sharing and group judgment. *Journal of Personality and social Psychology*, 65(5):959.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Graefe, A., Armstrong, J. S., Jones Jr, R. J., and Cuzán, A. G. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting*, 30(1):43–54.

Kim, O., Lim, S. C., and Shaw, K. W. (2001). The inefficiency of the mean analyst forecast as a summary forecast of earnings. *Journal of Accounting Research*, 39(2):329–335.

Lichtendahl Jr, K. C., Grushka-Cockayne, Y., and Pfeifer, P. E. (2013). The wisdom of competitive crowds. *Operations Research*, 61(6):1383–1398.

Lichtendahl Jr, K. C. and Winkler, R. L. (2007). Probability elicitation, scoring rules, and competition among forecasters. *Management Science*, 53(11):1745–1755.

Makridakis, S. and Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management science*, 29(9):987–996.

Mannes, A. E., Larrick, R. P., and Soll, J. B. (2012). The social psychology of the wisdom of crowds.

Mannes, A. E., Soll, J. B., and Larrick, R. P. (2014). The wisdom of select crowds. *Journal of personality and social psychology*, 107(2):276.

Martinie, M., Wilkening, T., and Howe, P. D. (2020). Using meta-predictions to identify experts in the crowd when past performance is unknown. *Plos one*, 15(4):e0232058.

Ottaviani, M. and Sørensen, P. N. (2006). The strategy of professional forecasting. *Journal of Financial Economics*, 81(2):441–466.

Palley, A. and Satopää, V. (2020). Boosting the wisdom of crowds within a single judgment problem: Selective averaging based on peer predictions. *Available at SSRN 3504286.*

Palley, A. B. and Soll, J. B. (2019). Extracting the wisdom of crowds when information is shared. *Management Science*, 65(5):2291–2309.

Pfeifer, P. E., Grushka-Cockayne, Y., and Lichtendahl Jr, K. C. (2014). The promise of prediction contests. *The American Statistician*, 68(4):264–270.

Prelec, D., Seung, H. S., and McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535.

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology*, 38(2):317–346.

Stekler, H. O., Sendor, D., and Verlander, R. (2010). Issues in sports forecasting. *International Journal of Forecasting*, 26(3):606–621.

Wilkening, T., Martinie, M., and Howe, P. D. (2021). Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems. *Management Science*.

Yaniv, I., Choshen-Hillel, S., and Milyavsky, M. (2009). Spurious consensus and opinion revision: Why might people be more confident in their less accurate judgments? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(2):558.

Appendices

A Proof of the theorem

Consider an expert judge $i \leq K$. Suppose all other experts and laypeople follow f^E and f^L respectively. Then,

$$E[\bar{x}_{-i}|s, t_i] = \frac{m + \ell(N-K)/(N-1)}{m + Kl} s + \frac{1}{N-1} \frac{N\ell}{m + K\ell} E\left[\sum_{j \neq i, j \in \{1,2,\ldots,K\}} t_j\right]$$

$$E[X|s, t_i] = \frac{m}{m + K\ell} s + \frac{\ell}{m + K\ell} \left(t_i + E\left[\sum_{j \neq i, j \in \{1,2,\ldots,K\}} t_j\right]\right)$$

The optimal report x_i^* satisfies

$$\frac{N-1}{N} E[\bar{x}_{-i}|s, t_i] + \frac{1}{N} x_i^* = E[X|s, t_i] \tag{8}$$

with expert i's expectations given above. Solving for x_i^* gives

$$x_i^* = f^E(s, t_i) = \frac{m - (N-K)\ell}{m + K\ell} s + \frac{N\ell}{m + K\ell} t_i$$

Thus, an expert judge i's best response is $f^E(s, t_i)$. Now, suppose judge i is a layperson instead, i.e. $i \in \{K+1, K+2, \ldots, N\}$. Then,

$$E[\bar{x}_{-i}|s] = \frac{m - \ell K/(N-1)}{m + K\ell} s + \frac{1}{N-1} \frac{N\ell}{m + K\ell} E\left[\sum_{j=1}^{K} t_j\right]$$

$$E[X|s] = \frac{m}{m + K\ell} s + \frac{\ell}{m + K\ell} E\left[\sum_{j=1}^{K} t_j\right]$$

The optimal report x_i^* satisfies the following condition:

$$\frac{N-1}{N} E[\bar{x}_{-i}|s] + \frac{1}{N} x_i^* = E[X|s]$$

which is the same condition as equation 8 except that a laypersons posterior expectations depend on s only. Solving for layperson i's x_i^* gives

$$x_i^* = f^L(s) = s$$

Thus, a layperson judge i's best response is $f^L(s)$. To summarize, $x_i^* = f^E(s, t_i)$ if $i \in \{1, 2, \ldots, K\}$ and $x_i^* = f^L(s)$ if $i \in \{K+1, K+2, \ldots, N\}$. Therefore, experts and laypeople following $f^E(s,t)$ and $f^L(s)$ respectively is an equilibrium. Furthermore, let γ be an experts' equilibrium relative weight on her private signal, given by

$$\gamma = \frac{N\ell/(m + K\ell)}{(m - (N-K)\ell)/(m + K\ell)} = \frac{N\ell}{m - (N-K)\ell}$$

An expert's relative weight on the private signal in her individual best guess (given in equation 2) is given by $\omega/(1-\omega) = \ell/m$. Observe that

$$\begin{aligned} \gamma &> \omega/(1-\omega) \\ \frac{N\ell}{m-(N-K)\ell} &> \frac{\ell}{m} \\ Nm &> m - (N-K)\ell \end{aligned}$$

which holds for all $N > 1$ and $K \leq N$. Thus, an expert i's equilibrium relative weight on t_i is strictly higher than the relative weight in her best guess.

B Summary statistics and additional figures

Table B1: Summary statistics, Study 1. The filtered sample excludes subjects who picked a wrong answer in the quiz (see the 'Procedure' in the main text) or picked 'Unclear' or 'Very Unclear' when asked for the clarity of the instructions.

	Experimental Condition			
	Individual	Crowd-5CG	Crowd-10CG	Crowd-30CG
Number of subjects	80	81	80	80
Female/Male	43/37	33/48	38/42	47/33
Average age	24.8	22.8	24	23.8
US/Non-US citizen	69/11	81/0	80/0	80/0
Average duration	5 min 14 sec	6 min	5 min 32 sec	5 min 13 sec
Average bonus	£1.04	£1.15	£1	£0.89
Number of subjects, filtered sample	73	75	72	72

Table B2: Summary statistics, Study 2. The filtered sample is constructed the same way as in Table B1

	Experimental Condition		
	Individual	Crowd-10	Contest-10
Number of subjects	84	128	81
Experts/Laypeople	-	81/47	-
Female/Male	36/48	33/48	38/42
Average age	23.4	24.6	23
US/Non-US citizen	72/12	103/25	65/16
Average duration	5 min 21 sec	5 min 35 sec	5 min 1 sec
Average bonus (Exp./Layp. in Crowd-10)	£1.26	£1.27/£0.49	£1.78
Number of subjects, filtered sample	69	113	64

Figure B1: The distribution of layperson predictions in Study 2.