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

Cem Peker †

Erasmus School of Economics, Erasmus University Rotterdam

September 2022

Abstract

Simple average of subjective forecasts is known to be effective in estimating uncertain quantities. However, benefits of averaging could be limited when forecasters 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 could induce self-extremization even in small crowds where winner-take-all contests (another incentive-based solution) are not effective.

*The author is grateful to Aurélien Baillon, Peter Wakker, the editor and two anonymous referees for their comments. This research was made possible by European Research Council Starting Grant 638408 Bayesian Markets.

†**Contact:** acpeker@gmail.com, https://orcid.org/0000-0001-9036-1915

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 & 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 multiple judgments typically produces accurate predictions (Armstrong, 2001). Aggregating 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. Optimal aggregation depends on the composition of the forecasting crowd (Lamberson & Page, 2012; Davis-Stober et al., 2015). Previous studies found simple averaging to be surprisingly effective and robust in a variety of estimation tasks (Genre et al., 2013; Clemen, 1989; Makridakis & 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 & Hastie, 1993; Chen et al., 2004). When best estimates of Bayesian experts are averaged, the shared information is over-represented in the aggregate prediction. As a result, the aggregate prediction exhibits the *shared-information bias* (Palley & 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 & Soll, 2019). Knowledge-weighting proposes a weighted combination of judgments (Palley & Satopää, 2022). The surprising overshoot (SO) algorithm picks a quantile from the empirical density of probability predictions (Peker, 2022). Pivoting, knowledge-weighting and the SO algorithm 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. Knowledge-weighting determines optimal weights based on the accuracy of meta-predictions. The SO algorithm infers the direction and size of the shared-information bias from the distribution of meta-predictions around the average prediction. Another line of work suggests weighting judgments according to judges' expertise in similar estimation tasks to improve the aggregate prediction (Budescu & 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 & Raftery, 2007). In contrast, we 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 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 to test if experts anticipate the shared-information problem and self-extremize in response. 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 the 'expert' role. These expert subjects observe an additional judge-specific sequence of sample flips, which represent their private signal. We construct forecasting crowds where each expert is grouped with multiple non-experts and rewarded for the accuracy of crowd average. The design makes the shared-information problem salient for experts as non-experts predictions are expected to be highly influenced by the shared signal. Evidence suggests that expert predictions are on average self-extremized 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 & Sørensen, 2006; Lichtendahl Jr & 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 experts do not significantly self-extremize under winner-take-all incentives in small crowds of experts. In contrast, incentives for crowd accuracy can elicit self-extremized predictions from a small number of experts in a large crowd.

Incentives for crowd accuracy encourage judges to consider their peers' judgments, and thus they may resemble beauty contest and guessing games (Camerer et al., 2004; Nagel, 1995). However, there are two important differences. Firstly, under incentives for crowd accuracy, rewards depend on the objective realization of an unknown quantity. So, the prediction task involves more than just anticipating others' judgments. Secondly, guessing games typically consider large samples where a single judge's report becomes negligible. Incentives for crowd accuracy consider finite samples in which a judge's prediction can influence the crowd average, which motivates self-extremization to improve accuracy.

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 & 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 & 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 & Soll (2019). In a linear aggregation problem,

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

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 & 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 $s \equiv (m_0\mu_0 + m_1 s_1)/m$. 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

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

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

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 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 & 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 judges 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.

Since we consider linear aggregation problems, we restrict our attention to reporting strategies of the form $f_E(s, t_i) = \alpha_1 s + \alpha_2 t_i$ and $f_L(s) = \beta s$ where f_E and f_L represent expert and layperson strategies respectively. The parameters $\{\alpha_1, \alpha_2, \beta\}$ denote the weights associated with reported predictions. Expert predictions can differ due to private signal t_i while laypeople report the same prediction given s. The case $\beta = 1$ corresponds to laypeople reporting their posterior expectation.

Definition. *An expert prediction is self-extremized if* $\alpha_2/(\alpha_1 + \alpha_2) > \omega$.

Recall that ω represents the weight on private signal in experts' individual best guess. Self-extremization is defined as the relative weight on private signal in the reported predictions being higher than ω. Note that we can have both $\alpha_1 > (1-\omega)$ and $\alpha_2 > \omega$ since α_1 and α_2 need not sum to unity. Thus, we describe self-extremization in terms of the normalized weight on private signal.

The theorem below presents an equilibrium of the simultaneous reporting game:

Theorem. *Under incentives for crowd accuracy, there exists infinitely many Bayesian Nash Equilibria such that*

$$x_i = \begin{cases} \alpha_1 s + \alpha_2 t_i & \text{for } i \in \{1, 2, \dots, K\} \\ \beta s & \text{for } i \in \{K+1, K+2, \dots, N\} \end{cases}$$

where $\{\alpha_1, \alpha_2, \beta\}$ *satisfy*

$$K\alpha_1 + (N-K)\beta = \frac{Nm}{m + K\ell} \tag{7}$$

$$\alpha_2 = \frac{N\ell}{m + K\ell} \tag{8}$$

$$\alpha_1, \alpha_2 \in \mathbb{R}, 0 < \beta \leq 1 \tag{9}$$

and experts self-extremize. For $K > 1$, self-extremization in expert judgments occurs for $\beta = 0$ as well.

Proof of the theorem is included in Appendix A. Conditions in 7 and 8 ensure that the resulting crowd average $\bar{x}$ does not exhibit the shared information bias. We have

$$\bar{x} = \frac{1}{N}\left\{\sum_{i=1}^{K} \alpha_1 s + \alpha_2 t_i + \sum_{i=1}^{K} t_i + \sum_{i=K+1}^{N} \beta s\right\} = \alpha_1 \frac{K}{N} s + \alpha_2 \frac{1}{N} \sum_{i=1}^{K} t_i + \beta \frac{N-K}{N} s$$
$$= \frac{m}{m+K\ell} s + \frac{\ell}{m+K\ell} \sum_{i=1}^{K} t_i$$

In equilibrium, the crowd average reflects the GPE given in equation 1. Experts and laypeople follow reporting strategies such that the shared and private signal are weighted optimally not in their individual predictions but in $\bar{x}$ instead. The decision maker does not need to select a subset of judges or determine weights for a weighted average. Simple averaging produces the optimal aggregate judgment.

The equilibria with $0 < \beta < 1$ represent situations where laypeople also coordinate on putting a lower weight on shared information. Experts self-extremize and the extent of their self-extremization depends on β. For $\beta = 0$, experts self-extremize for $K \geq 2$ even though laypeople put zero weight on the shared signal. The case $K = 1$ is the exception where single expert's optimal relative weight on t_i corresponds to ω in her posterior. Thus, the expert prediction is not self-extremized according to the definition above. However, the expert puts a higher absolute weight on both signals. Finally, we have the following equilibrium:

Corollary. *In the Bayesian Nash equilibrium with $\beta = 1$, laypeople simply report their posterior and experts self-extremize such that $\bar{x}$ does not exhibit the shared-information bias.*

The theorem characterizes type-symmetric equilibria in pure strategies with linear reporting. There exists many coordination equilibria where judges of the same type follow different strategies. Thus, only a subgroup of experts may self-extremize. Furthermore, the theorem characterizes equilibria with $\beta \in [0, 1]$. In a strategy with $\beta < 0$, laypeople put a negative weight on shared signal. Sufficient negative weighting from laypeople could correct the shared-information bias in $\bar{x}$ without self-extremization from experts. We may consider the equilibrium in the corollary ($\beta = 1$) most relevant, mainly because laypeople simply report their posterior. The theorem assumes common knowledge of information structure and composition of the forecasting crowd (i.e. values of K and N). Experts and laypeople coordinate on setting $\{\alpha_1, \alpha_2, \beta\}$ given their knowledge of $\{\ell, m, K, N\}$. In practice, only experts may have the knowledge that would allow them to anticipate the shared-information problem. If experts know the information structure and $\{\ell, m, K, N\}$, we could still observe the equilibrium outcome with $\beta = 1$, corresponding self-extremization in expert predictions, and no shared-information bias in $\bar{x}$.

Lichtendahl Jr et al. (2013) establish a limiting equilibrium in a Normal model where winner-take-all contests elicit self-extremized expert predictions in large crowds of experts. Note that for $K = N$ and $N \to \infty$, the optimal weight on private signals is 1 for any $\ell > 0$ and we have $\alpha_2 \to 1$ in the equilibrium above. Lichtendahl Jr et al. (2013) also show that, depending on the parameters, the limiting weight on the private signal is 1 either in a symmetric pure strategy equilibrium or in a mixed strategy equilibrium where experts provide a noisy report of their private signal only. These equilibria achieve optimal weighting of signals for $N \to \infty$. However, note that the equilibria in winner-take-all contests are limiting: the shared-information bias is alleviated only in large crowds. Incentives for crowd accuracy achieve optimal aggregation for any finite N and $K \leq N$ as well as the limiting case.

Since experts are the only source of private information, optimal weighting of private signals in $\bar{x}$ rely on expert predictions. Incentives for crowd accuracy would not work unless the experts anticipate the shared-information problem in $\bar{x}$ and self-extremize accordingly. Section 4 presents preliminary evidence from two experimental studies. Subjects are asked to predict the number of heads in 100 flips of a biased coin. Prior to making a prediction, subjects in the expert role observe shared and private signals, which consist of independent sequences of sample flips. We implement incentives for crowd accuracy to investigate if self-extremization occurs.

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.[2] In both studies the experimental design is similar to studies 1 and 2 in Palley & 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 test if experts self-extremize when 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

[2]Supplemental material includes the IRB approval for both studies granted by ERIM Internal Review Board, Section Experiments. The approval is registered under nr 2020/11/18-65868ape.

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 inclusion of CG agents makes the shared-information problem recognizable for subjects. Thus, Study 1 offers preliminary evidence on whether experts can anticipate the necessity of self-extremization. We implement a control group where subjects in expert role are rewarded for their individual accuracy and test if subjects self-extremize in the treatment conditions. Furthermore, we investigate if the crowd size has an impact on the rate of self-extremization. In small crowds, subjects may not perceive the severity of the shared-information problem and self-extremize less often. 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 diminish incentives to self-extremize. 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 - Do experts self-extremize?

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 within $[0.25, 0.75]$ and drawn uniformly. 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 a linear aggregation problem with a binomial variable (Palley & 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. Unlike in the theoretical framework, subjects' predictions are bounded within $[0, 100]$. The effect of censoring on reports will be discussed in Section 4.2.2.

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.

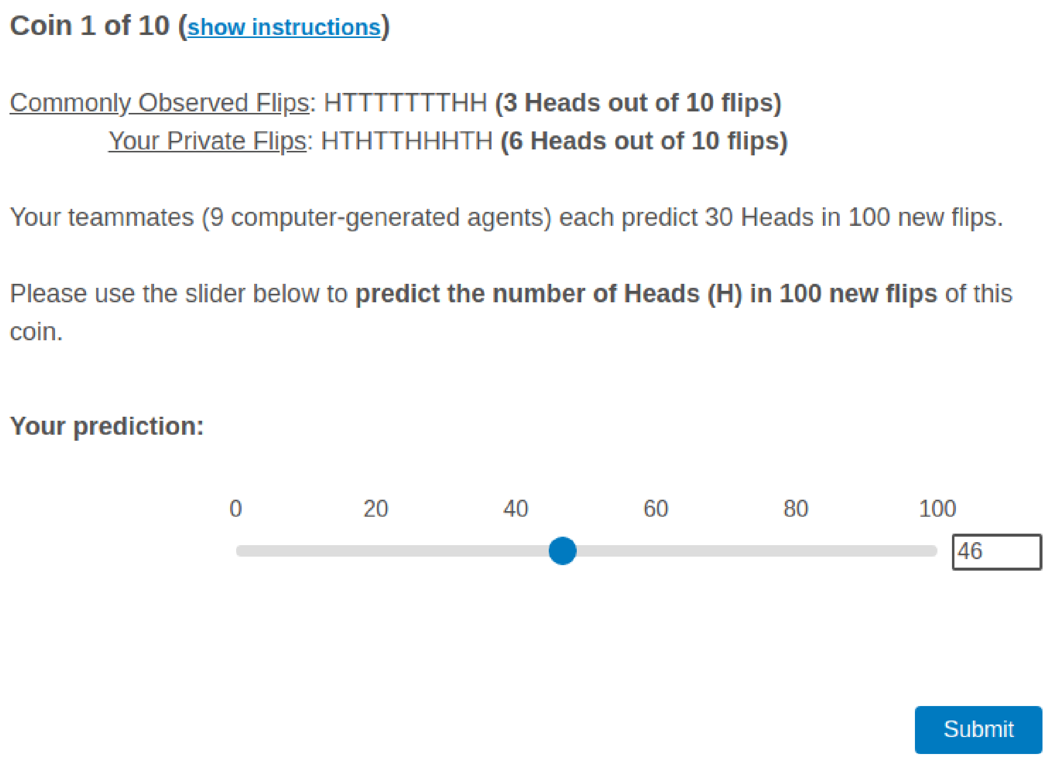

Figure 1: An example prediction task in the Crowd-10 condition. Initially, the slider starts at 0 and the text box that shows the current value is empty. The interface requires subjects to move (and release) the slider at least once or type a value directly.

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

$$B(x_i, x) = \begin{cases} 3 - \dfrac{1}{27}(x_i - x)^2 & \text{for } |x_i - x| \leq 9 \\ 0 & \text{for } |x_i - x| > 9 \end{cases} \tag{10}$$

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 $3 - (1/27)5^2 = £2.07$. 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. Figure C1 in Appendix C shows how bonuses are communicated to the subjects.

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.[3] 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. The experimental setup allows a precise definition 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

[3] Supplemental material provides all experimental data, instructions, quiz screens and the R Scripts (R Core Team, 2020; Wickham et al., 2022; Wickham, 2016, 2007; Leifeld, 2013) for reproducing all the empirical results.

subject's posterior best guess is 45. This subject's prediction is identified as self-extremized if it is higher than 45. If the reported prediction is lower than 45 instead, it would be considered as anti-extremized. Heterogeneity across individuals and reporting errors may lead to anti-extremized predictions as well as self-extremization in all treatments. However, if incentives for crowd accuracy motivate self-extremization, we should observe a higher percentage of extremized predictions in Crowd-5CG, Crowd-10CG and Crowd-30CG and similar rates of anti-extremization across all experimental conditions. Figure 2 shows the self-extremization rate in each experimental condition for various values of absolute difference between subjects' shared and private signals. Error bars indicate bootstrap standard errors.

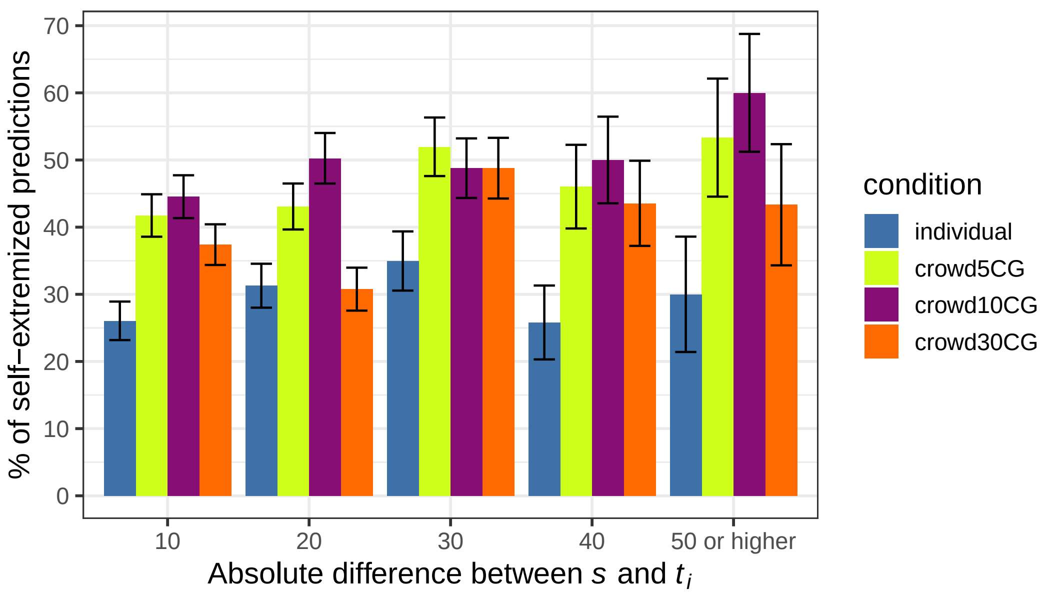

Figure 2: Self-extremization rate as measured by percentage of self-extremized predictions. Error bars show bootstrap standard errors (1000 boostrap samples).

Figure 2 indicates significantly higher self-extremization rate in crowd accuracy conditions, even when the shared and private signals are close and an expert would expect a small shared-information bias in the crowd average. Subjects anticipate the shared-information problem and adjust their prediction away from the shared signal. Figure 3 depicts the frequency of predictions that are equal to the posterior, extremized and anti-extremized. Figure 3 shows a substantially higher extremization rate in crowd accuracy conditions while the frequency of anti-extremized predictions is similar across all treatments. Subjects are more likely to adjust their predictions away from their posterior under incentives for crowd accuracy. Figure 3 suggests that the adjustments are in the direction of self-extremization.

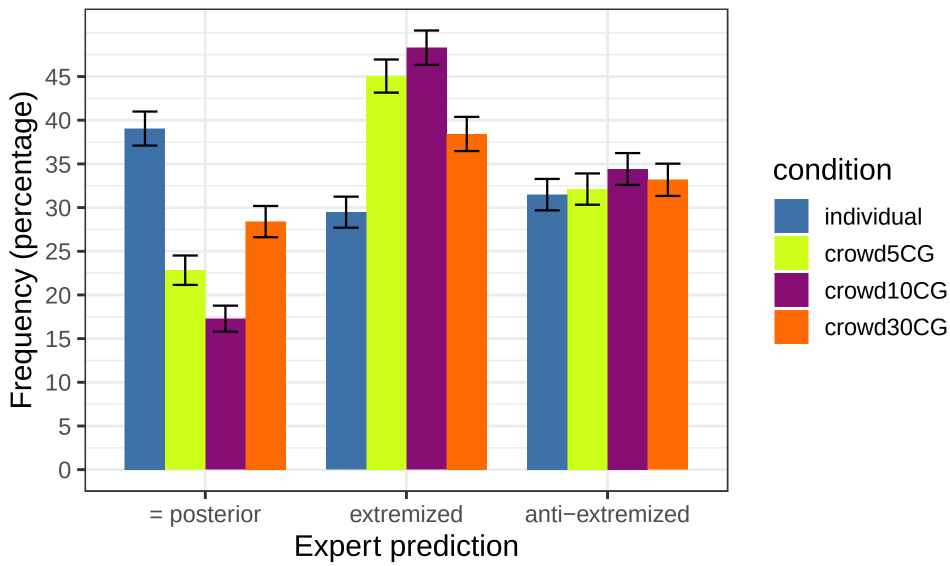

Figure 3: Frequency distribution of extremized and anti-extremized predictions in Study 1. Error bars show bootstrap standard errors (1000 boostrap samples).

Another variable of interest is the extent of self-extremization. Consider again the example in Figure 1 where the shared and private signals are 30 and 60 respectively and the posterior is 45. Suppose subject i reported $x_i = 50$. We refer to $50 - 45 = 5$ as the *extremizing adjustment*. In this example, the extremizing adjustment would be negative if the subject's report were less than 45. Positive and negative extremizing adjustments correspond to extremized and anti-exremized predictions respectively. We investigate if the extremizing adjustments of subjects who self-extremized are as extensive as predicted by the theory. For example, consider a subject in the Crowd-5CG who observed 6 and 7 heads in shared and private flips respectively. This subject's posterior is 65 but her optimal report (based on the theorem) is 85. So, the optimal extremizing adjustment is 20. Note that predictions in our task are bounded in $[0, 100]$ and the optimal prediction need not fall in that interval. For example, the optimal prediction in Figure 1 is 180 while subjects can self-extremize up to

100 only. In such tasks, we consider the maximum possible extremization as the optimal since extremizing as much as possible is expected to improve accuracy. In the case of Figure 1, the induced posterior is 45 and we consider $100 - 45 = 55$ as the optimal extremizing adjustment, which occurs if the subject reports 100.

For an analysis on the extent of self-extremization, we calculate extremizing adjustments as a percentage of the optimal. If the optimal adjustment is 20, an extremizing adjustment of 10 would be 50% of the optimal. Figure 4 depicts the frequency of percentage extremizing adjustments. Black bars represent predictions that are not self-extremized, i.e. extremizing adjustment is 0 or negative. Color-coded segments show self-extremized predictions where each color represent a range of extremizing adjustments as a percentage of the corresponding optimal adjustment.

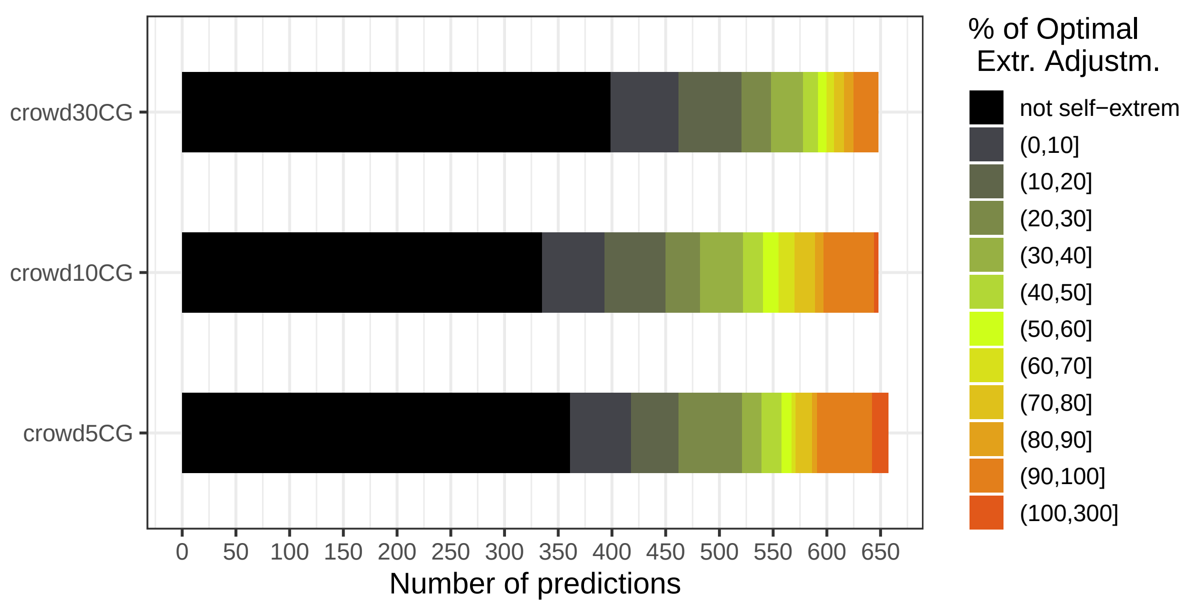

Figure 4: Extremizing adjustments as percentage of the corresponding optimal extremizing adjustment. Black bars represent predictions that are not self-extremized. Color-coded segments show the number of instances where the extremizing adjustment in 'percentage of the optimal' terms falls within the indicated interval.

In all three conditions, most extremizing adjustments fall short of the optimal. There are cases of excessive self-extremization as well. However, note that censoring in predictions affect the measurement of excessive self-extremization, in particular in Crowd-10CG and

Crowd-30CG. The optimal adjustment typically corresponds to reporting 0 or 100. Thus, extremizing adjustment cannot be higher than the optimal adjustment itself. Censoring could also be an explanation for slightly lower self-extremization rate in the Crowd-30CG condition in Figure 2. Subjects may reason that they cannot extremize enough to make a sizeable difference in accuracy, which would diminish the motivation to self-extremize. Figure C3 in Appendix C depicts the average extremizing adjustments at the subject level. Average adjustments are typically small in quantity and negative for some subjects. We observe that few subjects consistently self-extremize at the level predicted by the type-symmetric equilibria in the Theorem. Evidence suggests substantial heterogeneity in expert behavior under incentives for collective accuracy. Section 5 provides further discussion on the practical limitations implied by these findings.

Table 1 below shows the estimates of the linear regression models where extremizing adjustment (including both positive and negative observations) 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 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 2 we can conclude that incentives for crowd accuracy induce self-extremization. Figure 4 showed that most extremizing adjustments are smaller than the optimal adjustment in the corresponding prediction task. Nevertheless,

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

results suggest that incentives for crowd accuracy could alleviate the shared-information problem.

The censoring in predictions may affect the estimates in Table 1. Subjects cannot self-extremize beyond 0 or 100, which could cause a downward bias in extremizing adjustments. Note that the estimated extremizing adjustment is significantly higher in crowd accuracy conditions than Individual despite the potential negative effect of censoring. This result can be interpreted as a strong indicator of self-extremization on average.

Section 4.1 argued that self-extremization may occur more often 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. Figure 2 suggested that subjects self-extremized more often in Crowd-10CG condition, but the bootstrap standard errors suggest no major difference. The estimated extremizing adjustment is highest for Crowd-10CG in Table 1. Pairwise tests of coefficients show no significant differences across the crowd accuracy conditions ($t = 0.97, p = 0.34$ in Crowd-10CG vs Crowd-5CG; $t = 1.28, p = 0.20$ in Crowd-10CG vs Crowd-30CG under model (1)). As discussed above, censoring may affect the estimated extremizing adjustments in particular for Crowd-10CG and Crowd-30CG.

The results of Study 1 indicate that incentives for crowd accuracy could 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.

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. This approach implements the equilibrum with $\beta = 1$, where laypeople report their posteriors and experts self-extremize. Rewarding layperson subjects for individual accuracy keeps 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 posteriors. Experts could still anticipate that laypeople predictions will reflect the shared information. Thus, we expect to observe self-extremization in expert predictions.

In Contest-10 condition, each subject is in the expert role and participates in a winner-take-all contest with 9 other subjects. 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. Crowd-10 relies on experts' ability to anticipate the shared-information bias and self-extremize to improve the accuracy of crowd average. Contest-10 is an implementation of a winner-take-all contest. An expert would like to incorporate shared information and report her best estimate to maximize her chances of winning the prize. However, the prize is split in the case of a tie. The distribution of predictions is likely to have a higher density around the shared information. An expert can reduce the possibility of a tie by extremizing away from the shared information. But, self-extremization could increase expected error and result in a lower chance of winning the prize. This trade-off determines the extent of self-extremization that maximizes the expected prize (Pfeifer et al., 2014). Ties are less likely in small samples, so the experts have an incentive to simply maximize their accuracy. Thus, we may not observe self-extremization in Contest-10. In contrast, we expect self-extremization in Crowd-10 based on the theorem and findings in Study 1.

Note that including laypeople in a winner-take-all contest does not make experts' incentives to self-extremize stronger. An expert's posterior best guess differs from a laypersons' as long as her private signal is different from the shared signal. So, experts who report their posterior do not expect a tie with laypeople predictions. Other experts who may have the same posterior creates an incentive to self-extremize. Contest-10 represents a symmetric setup where winner-take-all incentives motivate self-extremization, except that the number of contestants is small.

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 10. 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-extremized 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

We analyze experts' predictions in each experimental condition. Figure C2 in Appendix C suggests that layperson subjects' predictions typically reflect the shared signal as in the equilibrium with $\beta = 1$. Figure 5 is analogous to Figure 3 and presents the frequency of extremized and anti-extremied predictions in Study 2.

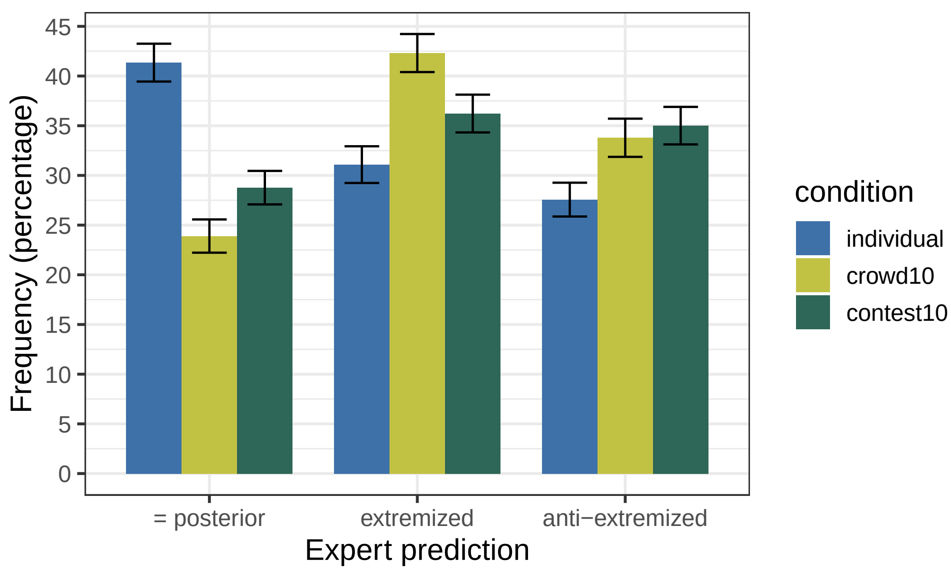

Figure 5: Frequency distribution of extremized and anti-extremized predictions in Study 2. Error bars show bootstrap standard errors (1000 boostrap samples).

Subjects deviate from their posterior more often in Crowd-10 and Contest-10 conditions. Percentage of extremized and anti-extremized predictions are similar in the Contest-10 condition. Subjects are almost as likely to put a higher weight on the shared signal and exacerbate the shared-information problem. In contrast, predictions that differ from the posterior are extremized more often in the Crowd-10 condition. Note that, unlike Figure 3, we also observe a higher frequency of anti-extremized predictions in the crowd accuracy condition. Recall that Study 1 made the shared-information problem highly salient. Subjects knew the exact predictions of computer-generated non-experts. The forecasting crowds in Crowd-10 consist of human subjects only. Expert subjects may find non-expert reports less predictable or incentives for crowd accuracy may lead to confusion for some individuals. Section 5 provides a discussion on the potential limitations of incentives for crowd accuracy.

Table 2 presents the regression estimates where extremizing adjustment is the dependent variable. 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.

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

Table 2 suggests a significantly higher level of extremizing adjustment in Crowd-10 than Individual. In contrast, there are no differences between the Contest-10 and Individual conditions in terms of extremizing adjustments. A pairwise comparison of Crowd-10 and Contest-10 also indicates a difference ($t = 2.61, p = 0.009$ in Crowd-10 vs Contest-10 under model (1)). Figure 5 showed that winner-take-all incentives in Contest-10 lead experts to deviate from their posteriors. However, extremizing adjustments are in the negative direction almost as often as the positive (self-extremizing) direction. As a result, estimated extremizing adjustment is not higher than the level observed in the Individual condition. Figure C4 in Appendix C depicts the distribution of experts' average extremizing adjustments in Study 2. Similar to Study 1, few experts systematically self-extremize across all tasks. Even though Table 2 indicates significant self-extremization on average, we should note the substantial heterogeneity in individual behavior where only a few subjects consistently self-extremize.

The results of Study 2 suggests that winner-take-all contests may not be effective if the forecasting crowd includes a small number of experts. Increasing the crowd size could help only if the decision maker can recruit more experts. Incentives for crowd accuracy could elicit self-extremized expert predictions in small crowds as well.

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 & Soll, 2019). In such cases, a decision maker would prefer experts to extremize their judgments away from the shared information. We propose incentivizing predictions 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 tested if experts follow the best response in the theorem given layperson predictions. Subjects are selected in forecasting crowds that consist of computer-generated non-experts with predictable predictions. Table 1 suggests that incentives for crowd accuracy generates self-extremization on average. However, we also observe substantial heterogeneity at the individual level. Most extremizing adjustments are less than optimal and only a small number of subjects self-extremized extensively in all tasks.

Study 2 tested incentives for crowd accuracy where experts are grouped with human non-expert subjects instead of computer-generated agents. 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 limiting equilibria 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. Table 2 indicates significant self-extremization under incentives for crowd accuracy but not in a winner-take-all contest. Figure 5 shows that subjects in the contest condition adjusted their forecast towards shared information almost as often as they self-extremized. Similar to Study 1, expert subjects' predictions under incentives for crowd accuracy exhibit considerable heterogeneity.

The influence of an individual prediction on the crowd average becomes smaller as the crowd size increases. Study 1 did not find significant differences in average extremizing adjustment across the crowd accuracy conditions. However, as discussed in Section 4.2.2, self-extremization in Crowd-10CG and Crowd-30CG may be affected by censoring in the experimental prediction task. Offering higher rewards for per unit reduction in the ex-post error of crowd average could make incentives to self-extremize stronger, in particular in large samples where a single judge's unit adjustment has a small impact on accuracy.

The coordination equilibrium in the theorem assumes common knowledge of the signal generation process and the composition of the forecasting crowd. The equilibrium outcome with $\beta = 1$ can still occur when non-experts lack such knowledge and simply report their posterior as long as experts coordinate on optimal self-extremization. The crowd accuracy conditions in our experimental studies circumvent the coordination problem by including a single expert only and focus on identifying if experts recognize the necessity of self-extremization. Study 1 simplifies the strategic considerations by using CG agents as laypeople. Study 2 incentivizes human laypeople subjects to report their posterior, which induces the equilibrium with $\beta = 1$. Figure 5 suggests that the presence of human laypeople leads to slightly higher rates of expert anti-extremization as well. We may expect further difficulties in coordination equilibria when there are multiple experts. Furthermore, the theorem considers only the type-symmetric equilibria while experimental evidence indicates substantial heterogeneity both at the prediction and individual levels. Non-symmetric equilibria could also be relevant for understanding expert behavior under incentives for crowd accuracy.

Incentives for crowd accuracy rely on Bayesian experts' ability to anticipate the shared-information problem. 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 & Yu, 2007; Yaniv et al., 2009). In our experimental studies, we grouped each expert subject exclusively with non-experts to make shared-information problem salient. Subjects had exact knowledge of the signal generation process and the number of laypeople in their crowd. Nevertheless, we observe considerable heterogeneity in expert behavior under incentives for crowd accuracy. The extent of extremization in self-extremized predictions is often less than optimal. Incentives for crowd accuracy may not induce sufficient self-extremization to fully correct for the shared-information bias. However, our experimental evidence suggests higher rates of self-extremization, which could alleviate the shared-information bias in the crowd average.

Presence of public knowledge could be the source of a salient shared-information problem in real-life forecasting tasks (Chen et al., 2004). Private information would reflect expert knowledge not accessible to laypeople. In mixed forecasting crowds, experts can anticipate that laypeople predictions rely exclusively on public knowledge. Subsequent empirical work may implement incentives for crowd accuracy in such prediction tasks and investigate if experts can coordinate on extremizing away from the shared information.

References

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

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

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

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, *119*(3), 861–898.

Chen, K.-Y., Fine, L. R., & 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., Broomell, S. B., & Dana, J. (2015). The composition of optimally wise crowds. *Decision Analysis*, *12*(3), 130–143.

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

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

Frongillo, R. M., Chen, Y., & Kash, I. A. (2015). Elicitation for aggregation. In *Twenty-ninth aaai conference on artificial intelligence.*

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

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

Gneiting, T., & 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., & Cuzán, A. G. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting*, *30*(1), 43–54.

Kim, O., Lim, S. C., & 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.

Lamberson, P., & Page, S. E. (2012). Optimal forecasting groups. *Management Science*, *58*(4), 805–810.

Leifeld, P. (2013). texreg: Conversion of statistical model output in R to LaTeX and HTML tables. *Journal of Statistical Software*, *55*(8), 1–24. Retrieved from `http://dx.doi.org/10.18637/jss.v055.i08`

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

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

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

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

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

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

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review*, *85*(5), 1313–1326.

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

Palley, A., & Satopää, V. (2022). Boosting the wisdom of crowds within a single judgment problem: Selective averaging based on peer predictions. Retrieved from `https://ssrn.com/abstract=3504286`

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

Peker, C. (2022). Extracting the collective wisdom in probabilistic judgments. *Theory and Decision*. doi: 10.1007/s11238-022-09899-4

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

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

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

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., & Verlander, R. (2010). Issues in sports forecasting. *International Journal of Forecasting*, *26*(3), 606–621.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2022). dplyr: A grammar of data manipulation [Computer software manual]. (https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr)

Wilkening, T., Martinie, M., & 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., & 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 $f_E(s, t_j) = \alpha_1 s + \alpha_2 t_j$ and $f_L(s) = \beta s$ respectively. Then,

$$E[\bar{x}_{-i}|s, t_i] = \frac{(K-1)\alpha_1 + (N-K)\beta}{N-1} s + \alpha_2 \frac{1}{N-1} E\left[\sum_{j \neq i, j \in \{1,2,\ldots,K\}} t_j \,\middle|\, s, t_i \right]$$

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

The optimal report x_i^* satisfies

$$\frac{N-1}{N} E[\bar{x}_{-i}|s, t_i] + \frac{1}{N} x_i^* = E[X|s, t_i] \tag{11}$$

with expert i's expectations given above. Plugging in we get

$$\frac{(K-1)\alpha_1 + (N-K)\beta}{N} s + \frac{1}{N} \alpha_2 E\left[\sum_{j \neq i, j \in \{1,2,\ldots,K\}} t_j \,\middle|\, s, t_i \right] + \frac{1}{N} x_i^* =$$
$$\frac{m}{m+K\ell} s + \frac{\ell}{m+K\ell} \left(t_i + E\left[\sum_{j \neq i, j \in \{1,2,\ldots,K\}} t_j \,\middle|\, s, t_i \right] \right)$$

Replace $K\alpha_1 + (N-K)\beta = Nm/(m+K\ell)$ and $\alpha_2/N = m/(m+K\ell)$ and solve for x_i^* to obtain

$$x_i^* = f_E(s, t_i) = \alpha_1 s + \alpha_2 t_i$$

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,

$$E[\bar{x}_{-i}|s] = \frac{\alpha_1 K + (N-K-1)\beta}{N-1} s + \alpha_2 \frac{1}{N-1} E\left[\sum_{j=1}^{K} t_j \,\middle|\, s\right]$$

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

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 11 except that a laypersons posterior expectations depend on s only. Plugging in the expectations we get:

$$\frac{\alpha_1 K + (N-K-1)\beta}{N} s + \alpha_2 \frac{1}{N} E\left[\sum_{j=1}^{K} t_j \,\middle|\, s\right] + \frac{1}{N} x_i^* = \frac{m}{m+K\ell} s + \frac{\ell}{m+K\ell} E\left[\sum_{j=1}^{K} t_j \,\middle|\, s\right]$$

Replace $\alpha_1 K + (N-K)\beta = Nm/(m+K\ell)$ and $\alpha_2/N = m/(m+K\ell)$ and solve for x_i^* to obtain

$$x_i^* = f_L(s) = \beta s$$

Thus, a layperson judge i's best response is $f_L(s)$. To summarize, $x_i^* = f_E(s, t_i) = \alpha_1 s + \alpha_2 t_i$ if $i \in \{1, 2, \ldots, K\}$ and $x_i^* = f_L(s) = \beta s$ 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, we have

$$\begin{aligned}
\frac{\alpha_2}{\alpha_1+\alpha_2} &= \frac{N\ell}{\frac{1}{K}(Nm-\beta(N-K)m)-\beta(N-K)\ell+N\ell} \\
&= \frac{NK\ell}{[N-\beta(N-K)](m+K\ell)}
\end{aligned}$$

Then we have

$$\begin{aligned}
\frac{\alpha_2}{\alpha_1+\alpha_2} &> \omega \\
\frac{NK\ell}{[N-\beta(N-K)](m+K\ell)} &> \frac{\ell}{m+\ell} \\
\frac{N}{N-\beta(N-K)} &> \frac{m+K\ell}{K(m+\ell)} \qquad (12)
\end{aligned}$$

Observe that for $\beta \in (0,1]$,

$$\frac{N}{N-\beta(N-K)} > 1 > \frac{m+K\ell}{K(m+\ell)}$$

for all $N > 1$ and $K \leq N$. Thus, experts self-extremize in equilibrium. Consider the case $\beta = 0$. Then, equation 12 is satisfied for $K > 1$, which implies experts self-extremize. For $K = 1$, the single expert's normalized weight is given by $N\ell/N(m+\ell) = \omega$.

B Summary statistics

	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 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-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

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

C Additional figures on design and results

Your bonus depends on the accuracy of your team's average. Here's an example:

Suppose there were 60 Heads in the 100 new flips. The table below shows the bonus for each value of your team's average:

Your team's average	Actual value	Your bonus
60	60	£3
59 or 61	60	£2.96
58 or 62	60	£2.85
57 or 63	60	£2.67
56 or 64	60	£2.41
55 or 65	60	£2.07
54 or 66	60	£1.67
53 or 67	60	£1.19
52 or 68	60	£0.63
51 or lower or 69 or higher	60	£0

Figure C1: How bonuses are displayed in the crowd accuracy conditions.

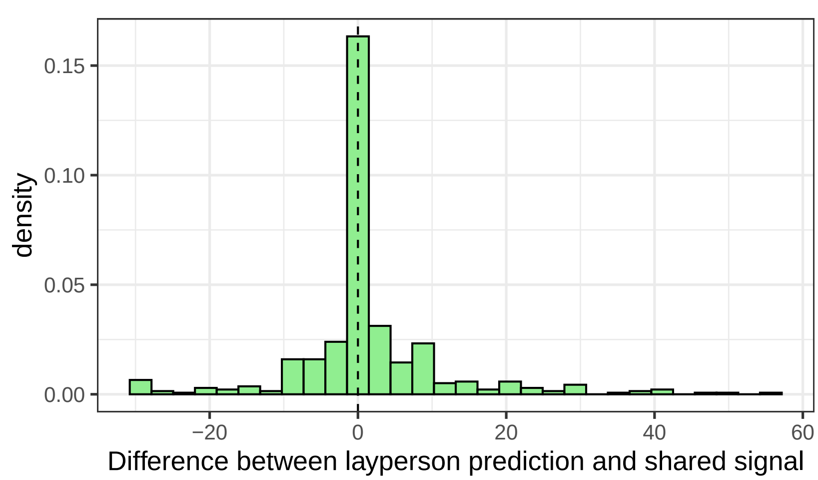

Figure C2: The distribution of layperson predictions in Study 2.

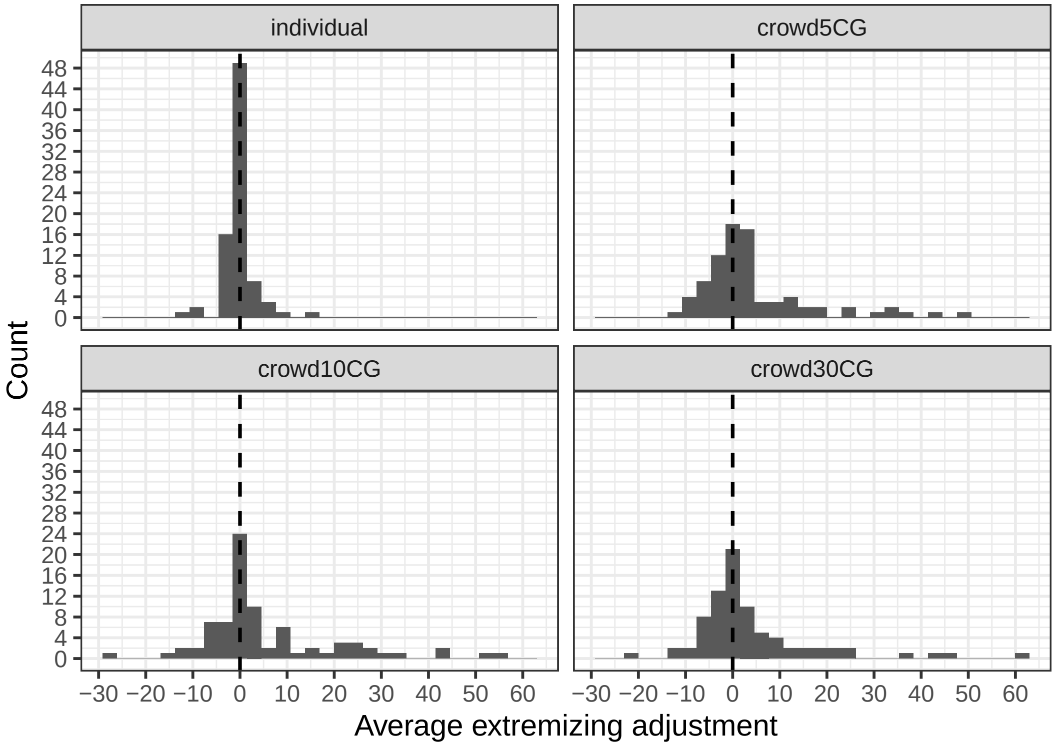

Figure C3: Average extremizing adjustments, Study 1

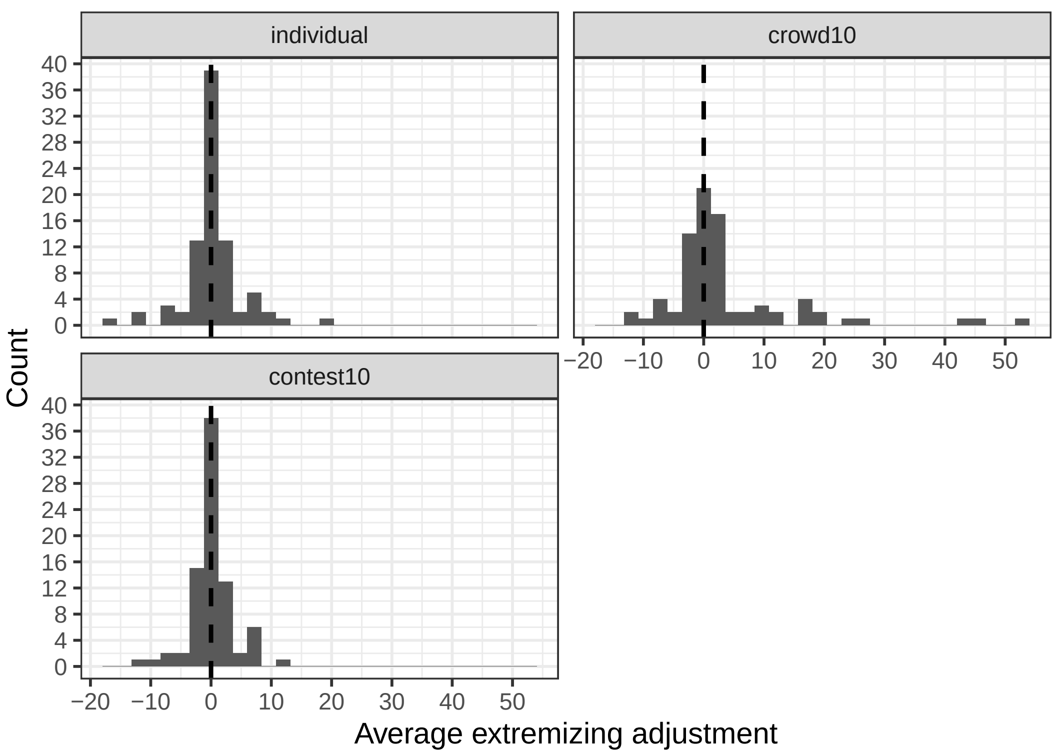

Figure C4: Average extremizing adjustments, Study 2