Extracting the collective wisdom of experts in probabilistic judgments*

Cem Peker†

December 2021

Abstract

How should we combine disagreeing expert judgments on the likelihood of an event? A common solution is simple averaging, which allows independent individual errors to cancel out. However, expert judgments can be correlated due to an overlap in their information, resulting in a miscalibration in the simple average. Optimal weights for weighted averaging are typically unknown and require past data to estimate reliably. This paper proposes an algorithm to aggregate probabilistic judgments under shared information. Experts are asked to report a prediction and a meta-prediction. The latter is an estimate of the average of other individuals' predictions. In a Bayesian setup, I show that if average prediction is a consistent estimator, the percentage of predictions and meta-predictions that overshoot the average prediction should be the same. An "overshoot surprise" occurs when the two measures differ. The Surprising Overshoot (SO) algorithm uses the information revealed in an overshoot surprise to produce a consistent estimator. Experimental evidence suggests that the algorithm performs well in moderate to large samples and in difficult aggregation problems where there is a strong disagreement between experts.

*I am grateful to Aurélien Baillon and Peter Wakker for their guidance and support. Special thanks to Tom Wilkening and Asa Palley for sharing their data and offering insights. Comments from the Behavioral Economics team at ESE and numerous seminar participants are much appreciated.

†Erasmus University Rotterdam, peker@ese.eur.nl

1 Introduction

In episode 'Errand of Mercy' of *Star Trek: The Original Series*, Captain Kirk and his first officer Mr. Spock find themselves in a dangerous situation. Captain Kirk plans to infiltrate an enemy camp and asks Mr. Spock the odds of getting out alive. Although being humanoid in appearance, Mr. Spock is an alien and his race is known for enhanced abilities in logical and quantitative reasoning. He estimates the odds to be '7824.7 to 1', which Captain Kirk half-jokingly calls 'a pretty close approximation'. In his usual seriousness, Mr. Spock responds by saying 'I endeavor to be accurate'.

The days of starships and deep space exploration may still be far away, but researchers and decision makers today find themselves facing uncertain decision problems, some of which may have stakes as high as Captain Kirk's. Scientists make probabilistic projections on natural phenomena, such as the occurrence of a major earthquake or the effects of anthropogenic climate change. Strategists assess the likelihood of important geopolitical events. Investors form judgments on the risks involved in investments. Economists and policy makers need probabilistic predictions on policy outcomes and macroeconomic indicators.

In such problems, the decision maker needs a well-calibrated probabilistic judgment. Unfortunately, a superior mind such as Mr. Spock's is not available (yet). Individual judgments may be subject to biases such as optimism, overconfidence, anchoring on an initial estimate, focusing too much on easily available information, neglecting an event's base rate, and many more (Kahneman and Tversky, 1973; Tversky and Kahneman, 1974; Kahneman et al., 1982). Combining multiple judgments to leverage 'the wisdom of crowds' is known to be an effective approach in improving accuracy (Surowiecki, 2004; Makridakis and Winkler, 1983).

The use of collective wisdom involves choosing an aggregation method that combines individual expert predictions into an aggregate prediction (Armstrong, 2001; Clemen, 1989; Palan et al., 2019). Previous work found simple averaging to be surprisingly effective, typically outperforming more sophisticated aggregation methods and showing robustness across various settings (Makridakis and Winkler, 1983; Mannes et al., 2012; Winkler et al., 2019;

Genre et al., 2013). Intuitively, simple averaging allows statistically independent individual errors to cancel, leading to a more accurate prediction (Larrick and Soll, 2006). However, in some prediction tasks, forecasters may have common information through shared expertise, past realizations, knowledge of the same academic works, etc. (Chen et al., 2004). Then, individual errors may become correlated, resulting in a bias in the equally weighted average of predictions (Palley and Soll, 2019). In theory, the decision maker in a given task can select and weight judgments such that the errors perfectly cancel out (Clemen and Winkler, 1986; Mannes et al., 2014; Budescu and Chen, 2015). However, optimal weights depend on how experts' prediction errors are correlated and are typically unknown to the decision maker. Conventional methods aim to estimate appropriate weights using past data from similar tasks (Budescu and Chen, 2015; Mannes et al., 2014). The effectiveness of this approach is limited by the availability and reliability of past data.

This paper develops an algorithm to aggregate judgments on the likelihood of an event. I consider a setup where experts form their judgments by combining shared and private information on the probability of an event. When the shared information differs from the true probability, experts are likely to err in the same direction, resulting in a miscalibrated average prediction. The algorithm relies on an augmented elicitation procedure commonly found in several recent papers (Prelec, 2004; Prelec et al., 2017; Palley and Soll, 2019; Palley and Satopää, 2020; Wilkening et al., 2021): Experts report a prediction of the probability as well as an estimate of the average of others' predictions, which is referred to as a meta-prediction. I show that when average prediction is a consistent estimator, the percentage of predictions and meta-predictions that overshoot the average prediction should be the same. Whenever the two measures differ an *overshoot surprise* occurs, which indicates that the average prediction is an inconsistent estimator of the unknown probability. I develop the Surprising Overshoot (SO) algorithm which produces a consistent estimator. The SO algorithm uses the information in the size and direction of the overshoot surprise. It does not require the use of past data.

I test the SO algorithm using experimental data from two sources. Palley and Soll (2019) conducted an experimental study where subjects are asked to predict the number of heads in 100 flips of a biased coin. The SO algorithm shares the same Bayesian framework as Palley and Soll (2019) and the experiment implements shared and private signals as sample flips from the biased coin. This study tests the SO algorithm in a controlled setup. The second source is Wilkening et al. (2021), who conducted two experimental studies. The first experiment replicates the earlier study by Prelec et al. (2017) which asked subjects true/false questions about the capital cities of U.S. states. However, unlike Prelec et al. (2017) they also ask subjects to report probabilistic predictions and meta-predictions, which allows an implementation of the SO algorithm. In the second experiment, Wilkening et al. (2021) generate 500 basic science statements and ask subjects to report probabilistic predictions and meta-predictions on the likelihood that a given statement is true. I use the data from these two experiments to investigate if the SO algorithm produces an aggregate probabilistic prediction closer to the correct answer. Results suggest that the SO algorithm outperforms simple benchmarks such as unweighted averaging and median prediction. I also compare the SO algorithm to alternative solutions for aggregating probabilistic judgments, which elicit similar information from individuals (Palley and Soll, 2019; Martinie et al., 2020; Palley and Satopää, 2020; Wilkening et al., 2021). The SO algorithm never underperforms and compares favorably to alternative aggregation mechanisms in prediction tasks where individual predictions are highly dispersed. Experimental evidence suggests that the SO algorithm is especially effective in extracting the collective wisdom from disagreeing probabilistic judgments in moderate to large samples of experts.

This paper contributes to the literature of judgment aggregation mechanisms that utilize meta-beliefs to improve prediction accuracy. The Surprisingly Popular algorithm picks an answer to a multiple choice question based on predicted and realized endorsement rates of alternative choices (Prelec et al., 2017). The Surprisingly Confident (SC) algorithm determines weights that leverage more informed judgments (Wilkening et al., 2021). The SP and

SC algorithms aim to find the correct answer to a binary or multiple-choice question while the SO algorithm produces a probabilistic estimate on a binary event.

Recent work developed aggregation algorithms for probabilistic judgments as well. Pivoting uses meta-predictions to recover and recombine shared and private information in the optimal way (Palley and Soll, 2019). Knowledge-weighting constructs a weighted crowd average (Palley and Satopää, 2020). Individual weights are estimated based on the accuracy of meta-predictions. The meta-probability weighting algorithm also attaches weights to individual predictions where the absolute difference between an individual's prediction and meta-prediction is considered as an indicator of expertise (Martinie et al., 2020). In testing the performance of the SO algorithm, pivoting, knowledge-weighting and meta-probability weighting are considered as benchmarks. As mentioned above, the SO algorithm performs especially well when individual judgments are highly dispersed. In practice, such problems are likely to be the most challenging ones, where expert judgments disagree substantially and it is not clear how judgments should be aggregated for maximum accuracy.

The rest of this paper is organized as follows: Section 2 introduces the formal framework. Section 3 develops the SO algorithm and establishes the theoretical properties of the SO estimator. Section 4 introduces the data sets and benchmarks we consider in testing the SO algorithm empirically. The same section also presents some preliminary evidence on how overshoot surprises relate to the inaccuracy in average prediction. Section 5 presents experimental evidence testing the SO algorithm. Section 6 provides a discussion on the effectiveness of the SO algorithm. Section 7 concludes.

2 The Framework

The framework follows the definition of a *linear aggregation problem* in Palley and Soll (2019) and Palley and Satopää (2020). Let $Y \in \{0, 1\}$ be a random variable that represents the occurrence of an event where $y \in \{0, 1\}$ denotes the value in a given realization. Also let $\theta = P(Y = 1)$ be the unknown objective probability of the outcome 1, representing the occurrence of the event. A decision maker (DM) would like to estimate θ. The DM elicits judgments from a sample of N agents to develop an estimator, where $N \to \infty$ represents the whole population. All agents observe a *shared signal* $s \in [0, 1]$. Each agent $i \in \{1, 2, \ldots, K\}$ where $K \leq N$ receives a *private signal* $t_i \in [0, 1]$. All signals follow the same expectation θ and are conditionally independent given θ. In the analysis below, we consider the case where $K = N$, i.e. all agents observe a private signal. Appendix B presents the same analysis for the case of $K < N$ and shows that the same results are applicable.

Each agent i follows a belief updating according to the Bayes' rule. Let $E[\theta|s, t_i]$ denote an agent i's posterior expectation of θ. In a linear aggregation problem, the posterior expectation combines the signals linearly:

$$E[\theta|s, t_i] = (1 - \omega)s + \omega t_i \tag{1}$$

where $\omega \in [0, 1]$ represents the weight an agent puts on the private signal relative to the shared signal s. It is common knowledge to the agents that the posterior expectation of any agent i is given by Equation 1. The weight parameter ω and the signals $\{s, t_1, t_2, \ldots, t_N\}$ are unknown to the decision maker. Palley and Soll (2019) specify such a linear aggregation problem. Agents share the common prior on θ given by a Beta density. The shared signal s and each agent's private signal t_i are given by the average of m and ℓ independent realizations of Y respectively. The signal structure and $\{m, \ell\}$ are common knowledge to all agents. Then, the posterior expectation of an agent i on θ is given by Equation 1 and $\omega = \ell/(m + \ell)$ represents the relative informativeness of t_i compared to s.

Suppose the DM considers the simple average of agents' predictions as an estimator for θ. Let x_i be agent i's reported prediction on θ. Suppose all agents report their best guesses, i.e. $x_i = E[\theta|s, t_i]$. Then the average prediction is given by

$$\bar{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_i = (1-\omega)s + \omega\frac{1}{N}\sum_{i=1}^{N} t_i.$$

Note that $\lim_{N\to\infty} \bar{x}_N = (1-\omega)s+\omega\theta \neq \theta$ if $s \neq \theta$, i.e. the average prediction is not a consistent estimator of θ unless the shared information is perfectly accurate (Palley and Soll, 2019). Increasing the sample size does not alleviate the shared-information problem as s is incorporated in $\bar{x}_N$ by each additional prediction. Shared information causes a correlation between predictions and leads to a persistent error in $\bar{x}_N$. Section 3 develops the Surprising Overshoot algorithm, which constructs an estimator that accounts for the shared-information problem and constructs a consistent estimator. Then, I will investigate the empirical effectiveness of the algorithm using experimental data from Palley and Soll (2019) and Wilkening et al. (2021).

3 The Surprising Overshoot algorithm

The Surprising Overshoot algorithm relies on an augmented elicitation procedure and the information revealed by the distribution of agents' reports to construct a consistent estimator. Section 3.1 introduces the elicitation procedure. Sections 3.2 and 3.3 elaborates on the relationship between agents' equilibrium reports and the resulting average prediction, which shows the feasibility of a consistent estimator. Section 3.4 develops the SO algorithm which produces the consistent estimator.

3.1 Elicitation procedure

The decision maker simultaneously and separately asks each agent i to submit two reports. In the first, the agent is asked to make a *prediction* on θ, denoted by x_i. In the second, the agent reports a *meta-prediction* $z_i \in \mathbb{R}$, which is an estimate of the average prediction of agents $j \in \{1, 2, \ldots, N\} \setminus \{i\}$, denoted by $\bar{x}_{-i} = \frac{1}{N-1} \sum_{j \neq i} x_j$. Agents' reports are incentivized by a strictly proper scoring rule (Gneiting and Raftery, 2007). Let $\pi_{xi} = S_x(x_i, y)$ and $\pi_{zi} = S_z(z_i, \bar{x}_{-i})$ be the ex-post payoffs of an agent i from the prediction and meta-prediction where S_x and S_z are strictly proper scoring rules satisfying $\theta = \arg\max_{u \in \mathbb{R}} S_x(u, Y)$ and $\bar{x}_{-i} = \arg\max_{u \in \mathbb{R}} S_z(u, \bar{x}_{-i})$. Agent i's total payoff is given by $\pi_i = \pi_{xi} + \pi_{zi}$.

An agent i's report is *truthful* if $(x_i, z_i) = (E[\theta|s, t_i], E[\bar{x}_{-i}|s, t_i])$, i.e. agent i reports her posterior expectations on θ and $\bar{x}_{-i}$ as prediction and meta-prediction respectively. Truthful reporting represents the situation where reports are truthful for all $i \in \{1, 2, \ldots, N\}$.

Theorem 1. *Truthful reporting is a Bayesian Nash equilibrium in the simultaneous reporting game*

All proofs are included in Appendix A. Intuitively, Theorem 1 follows from the use of proper scoring rules. Agents are incentivized to report their best estimates on the unknown probability and the average of others' predictions. In equilibrium, we have $x_i = E[\theta|s, t_i] = (1-\omega)s + \omega t_i$ for all $i \in \{1, 2, \ldots, N\}$. Then, agent i's equilibrium meta-prediction is given by $E[\bar{x}_{-i}|s, t_i] = (1-\omega)s + \omega \frac{1}{N-1} \sum_{j \neq i} E[t_j|s, t_i]$. Observe that $E[t_j|s, t_i] = E[E[t_j|\theta]|s, t_i] = E[\theta|s, t_i]$, i.e. agent i's expectation on another agent's signal is her expectation on θ, which is equal to the truthful prediction. Thus, the equilibrium prediction and meta-prediction of an agent i are given by:

$$x_i = (1-\omega)s + \omega t_i \tag{2}$$

$$z_i = (1-\omega)s + \omega x_i \tag{3}$$

In the remainder of this section, I assume truthful reporting and hence, each agent i's reported predictions and meta-predictions are given by equations 2 and 3 respectively.

3.2 Overshoot rates in predictions and meta-predictions

A prediction or meta-prediction is said to *overshoot* the average prediction $\bar{x}_N$ if it exceeds $\bar{x}_N$. For any arbitrary agent i, there are two overshoot indicators. For example, if $x_i > \bar{x}_N > z_i$, agent i's prediction x_i1 overshoots the average prediction while the meta-prediction z_i does not overshoot.

Suppose $x_i > \bar{x}_N$ for an agent i. For this agent, we can write

$$\begin{aligned} x_i &> \bar{x}_N \\ (1-\omega)s + \omega t_i &> (1-\omega)s + \omega \frac{1}{N}\sum_{k=1}^{N} t_k \\ t_i &> \frac{1}{N}\sum_{k=1}^{N} t_k = \bar{t} \end{aligned}$$

where $\bar{t}$ denotes the average private signal. Recall that θ is the mean of the signal distribution. We have $\bar{t} \to \theta$ for $N \to \infty$. In the limit, any given agent i's prediction is higher than the average prediction if and only if agent i received a higher than average signal, that is

$$x_i > \bar{x} \iff t_i > \theta \tag{4}$$

where $\bar{x} = \lim\limits_{N\to\infty} \bar{x}_N$ is the population average of predictions.

Now suppose $z_j > \bar{x}_N$ for an agent j (who may or may not be the same agent as i above). The following holds for z_j:

$$\begin{aligned} z_j &> \bar{x}_N \\ (1-\omega)s + \omega x_j &> (1-\omega)s + \omega \frac{1}{N}\sum_{i=k}^{N} t_k \\ x_j &> \frac{1}{N}\sum_{k=1}^{N} t_k = \theta \quad \text{for} \quad N \to \infty \end{aligned}$$

Then, for $N \to \infty$ we get

$$z_j > \bar{x} \iff x_j > \theta \tag{5}$$

All quantities in equation 5 except θ are known to the DM, which has an important implication about the identifiability of θ for $N \to \infty$. Equation 5 implies $z_j = \bar{x} \iff x_j = \theta$. So, if agent j's meta-prediction z_j equals the population average of predictions, the prediction x_j reflects the unknown probability θ. This observation motivates the Surprising Overshoot algorithm introduced below. Furthermore, equations 4 and 5 together suggest a pattern for $N \to \infty$. In equation 4, an agent i's prediction x_i overshoots $\bar{x}$ if and only if $t_i > \theta$. However, for meta-prediction z_i to overshoot $\bar{x}$, we must have $x_i = (1 - \omega)s + \omega t_i > \theta$. Thus, we do not necessarily have $z_i > \bar{x}_i$ whenever $x_i > \bar{x}$ is satisfied. Consider the following measures computed using predictions and meta-predictions:

$$p_x = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(x_i > \bar{x})$$
$$p_z = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(z_i > \bar{x})$$

The measures p_x and p_z represent the population proportion of predictions and meta-predictions that overshoot the population mean $\bar{x}$. I refer to p_x and p_z as the *overshoot rate* in predictions and meta-predictions respectively. From equation 5, we can infer that p_z also corresponds the population proportion of predictions such that $x_i > \theta$.

3.3 Overshoot surprise as an indicator of the inconsistency in the simple average of predictions

A comparison between the overshoot rates in predictions and meta-predictions reveals a potential inconsistency in the average prediction. Let $f(x)$ denote the population density of predictions given the signal distribution. Figure 1 illustrates two scenarios with an example distribution:

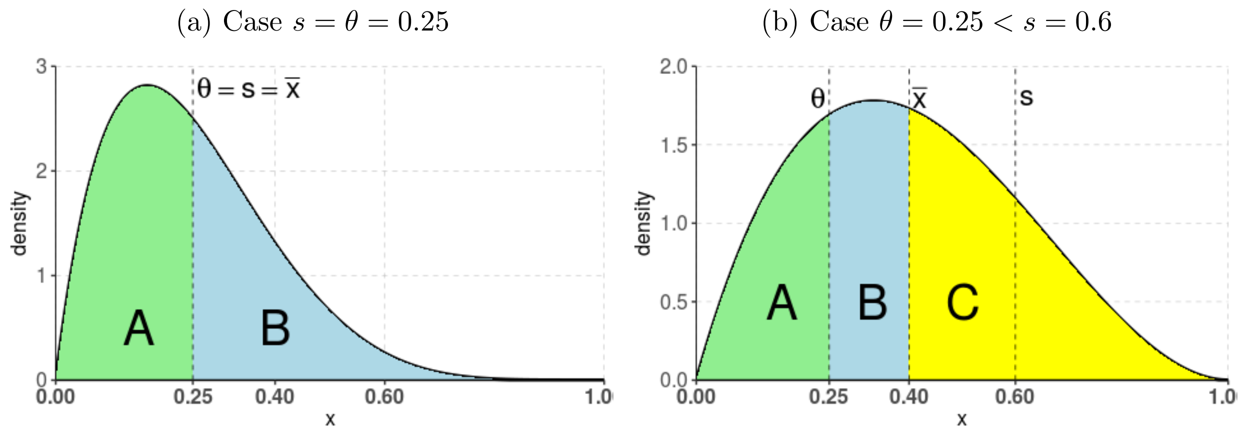


Figure 1: An example theoretical density f of forecasts in two example cases. A, B and C represent the shaded regions.

Figure 1a represents the case where $\bar{x}$ and θ coincide, which follows from $s = \theta$. Then, p_x and p_z both represent the region B and we have $p_x = p_z$. The average prediction $\bar{x}$ reflects θ. In contrast, Figure 1b shows an example density of predictions where the shared signal s differs from θ, resulting in $\bar{x} > \theta$. Observe that in Figure 1b, p_z represents the region $B + C$ while p_x corresponds to the area marked by C only. In this case, we see $\bar{x} > \theta$ due to $s > \theta$. As a result, the percentage of predictions that exceed $\bar{x}$ is lower and we have $p_z > p_x$.

Definition 1 (**Overshoot surprise**). *An overshoot surprise occurs when $p_x \neq p_z$. The overshoot surprise is positive if $p_z > p_x$ and negative if $p_z < p_x$. The size of the overshoot surprise is given by $\Delta p = p_z - p_x$.*

In Figure 1b, we observe a positive overshoot surprise, which indicates that $\bar{x} > \theta$. A negative overshoot surprise would ensue if $\bar{x} < \theta$ instead. In contrast, there is no overshoot surprise in Figure 1a where $\bar{x}$ is perfectly accurate. Furthermore, the size of the overshoot surprise is related to the absolute difference between the shared signal s and θ. When $|s - \theta|$ is higher, $\bar{x}$ will be further away from θ, resulting in a bigger overshoot surprise.

3.4 The Surprising Overshoot estimator and its consistency

Equation 5 and the example in Figure 1 suggest that p_z can be used to recover θ. Consider the cumulative density F associated with the density of predictions in Figure 1b:

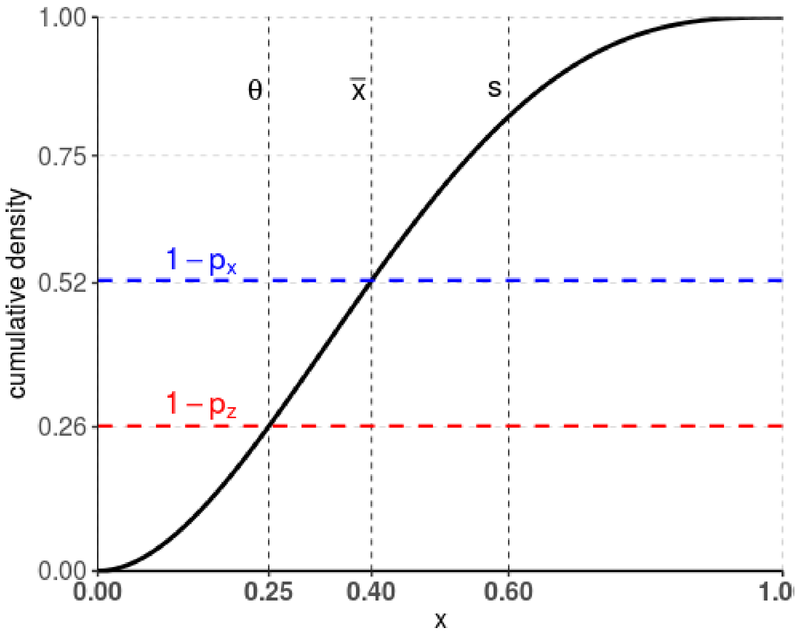


Figure 2: Cumulative density F of predictions for the example in Figure 1b

Since p_z measures the proportion of predictions that overshoot θ, $1-p_z$ gives the quantile of the cumulative density that corresponds to θ. We can generalize this finding as follows: Let F denote the continuous population density of predictions. Then, $\theta = F^{-1}(1-p_z)$. Lemma 1 suggests that when the predictions and meta-predictions in the whole population ($N \to \infty$) are known and the population density of predictions is continuous, the exact θ can be recovered using p_z and F. The measure p_z simply reveals the cumulative density of predictions that overshoot θ, implying that $1-p_z$ gives the quantile of F that corresponds to θ.

In practice, a DM can only recruit a finite sample of agents. The population distribution of predictions and the population quantiles are unknown. Let $\hat{F}_N$ be the empirical cumulative density of the predictions in a finite sample of agents of size N. I maintain the assumption that F is a continuous cumulative density. Also let $\hat{p}_{zN} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(z_i > \bar{x}_N)$ be the sample overshoot rate in meta-predictions in a finite sample of size N. The definition below introduces the Surprising Overshoot (SO) algorithm:

Definition 2 (**The Surprising Overshoot algorithm**). *The Surprising Overshoot algorithm constructs the estimator x_N^{SO} for θ following the steps below:*

1. *Elicit sample predictions and meta-predictions in a finite sample of size N.*

2. *Calculate $\hat{p}_{zN} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(z_i > \bar{x}_N)$.*

3. *Set $x_N^{SO} = inf\{y|\hat{F}_N(y) \geq 1 - \hat{p}_{zN}\})$.*

The SO algorithm simply locates the $1 - \hat{p}_{zN}$ quantile of the sample predictions. Lemma 1 established that when predictions follow a continuous density, the unknown probability is the $1 - p_z$ population quantile. Since the population density and p_z are unknown, the SO algorithm relies on the sample counterparts.

Theorem 2. *The SO estimator x_N^{SO} is a consistent estimator of θ.*

The SO estimator may exhibit a non-zero bias in finite samples. However, unlike the average prediction, the SO estimator converges to θ in the limit. Thus, we can expect x_N^{SO} to achieve higher accuracy than $\bar{x}_N$ as we increase N.

Intuitively, the SO estimator can be considered as the average prediction adjusted in a particular direction. A positive (negative) overshoot surprise indicates that $\bar{x}_N$ overestimates (underestimates) θ even for $N \to \infty$. Then, x_N^{SO} is expected to be lower (higher) than $\bar{x}_N$ as $1 - \hat{p}_{zN}$ is expected to be a smaller (larger) quantile than $\bar{x}_N$. Section 4 presents empirical evidence suggesting that overshoot surprises strongly correlate with the forecasting errors of average prediction as predicted by the theoretical framework. Thus, the SO estimator adjusts the average prediction in the correct direction. When there is no overshoot surprise, we expect the SO algorithm to produce $x_N^{SO} = \bar{x}_N$ and not distort the consistent estimator.

The SO estimator relies on the empirical distribution of predictions as well as agents' meta-predictions. This property has implications about the prediction problems where we may expect the SO algorithm to be more effective. To illustrate, consider the two example empirical densities below. Both figures depict predictions from a sample of 10 agents where the sample average prediction is 0.4 while $\theta = 0.25$. In Figure 3a agents report one of $0.5, 0.3$ or 0.1 as prediction. The distribution of predictions in Figure 3b is more dispersed around the average prediction. Suppose the meta-predictions in each example (not shown on figures) are such that $\hat{p}_{zN} = 0.2$ in both cases. Then the SO estimate is $1 - \hat{p}_{zN} = 0.8$ quantile of the empirical density of predictions. The orange bar in each figure locates the SO estimate.

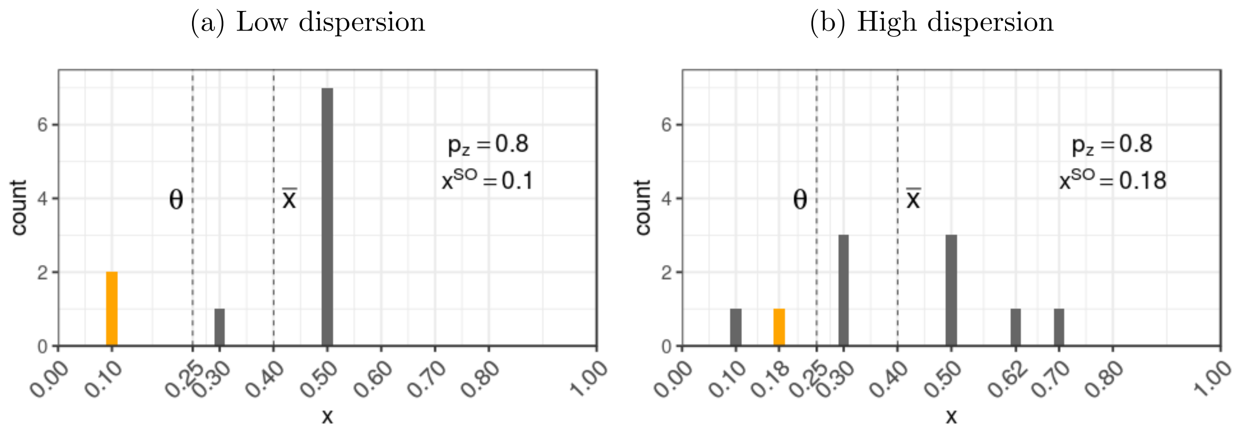


Figure 3: Two examples of empirical density of predictions

The SO estimate is more accurate in the high dispersion case simply because the 0.2 quantile falls closer to θ. The SO algorithm picks the prediction that corresponds to the sample quantile $1 - \hat{p}_{zN}$. So the set of values x_N^{SO} can take depends on the empirical density of predictions. Even when $1 - \hat{p}_{zN}$ provides an accurate estimate of the cumulative density at θ, the SO estimate may not be more accurate than $\bar{x}_N$ simply because $1 - \hat{p}_{zN}$ quantile of the sample predictions is not close to θ. Such cases are less likely when the sample size is higher and/or the empirical density of predictions is more dispersed, as in Figure 3b. Therefore, we may expect the SO algorithm to perform better in larger samples and when the predictions are more dispersed. Intuitively, high dispersion can be considered as representing prediction tasks where individual judgments disagree, which could occur when the event of interest is highly uncertain and there is no strong expert consensus. The following sections test the

SO algorithm using experimental data. In the analyses below, sample size and dispersion of predictions are considered to be the factors of interest.

4 Testing the SO algorithm

This section outlines the empirical methodology and presents some preliminary evidence on overshoot surprises. I use data from various experimental studies to test the SO algorithm. Section 4.1 provides information on the data sets. In testing the SO algorithm, I follow a comparative approach. The analysis will implement various alternative methods as a benchmark and test if the SO algorithm performs significantly better. Section 4.2 introduces the benchmarks. Finally, Section 4.3 provides some preliminary findings on overshoot surprises and how they relate to the inconsistency in the simple average of predictions.

4.1 Data sets

I use data from three experimental studies. The first data set comes from Study 1 in Palley and Soll (2019). They conducted an online experiment where subjects reported their prediction and meta-prediction on the number of heads in 100 flips of a biased two-sided coin. The actual probability of heads is unknown to the subjects. Prior to submitting a report on a coin, each subject observed two independent samples of flips. One sample is common to all subjects and represents the shared signal. The second sample is subject-specific and constitutes a subject's private signal. A subject's best guess on the number of heads in 100 new flips a coin is effectively that subject's best guess on the unknown bias. Thus, the prediction task elicits predictions about a probability unknown to subjects.

Study 1 in Palley and Soll (2019) implements 3 different information structures. All subjects observe the shared signal and a private signal in the 'Symmetric' setup while only a subset of subjects observe a private signal in the 'Nested-Symmetric' structure. Private signals are subject-specific and unbiased in both structures, which agrees with the theoretical framework of the SO algorithm. The other setup is referred to as the 'Nested' structure, in which private signals are not subject-specific. The average of private signals do not converge to the true value, which deviates from the theoretical framework of the SO algorithm. Thus, I exclude the 'Nested' structure and use the 48 prediction tasks (48 distinct coins) from the 'Symmetric' and 'Nested-Symmetric' structures only.

The Coin Flips data set from Palley and Soll (2019)'s Study 1 allows testing the SO algorithm in a controlled setup. Since the unknown probabilities are known to the analyst, it is possible to calculate prediction errors directly. The number of subjects per coin vary between 101 and 125, which allows an analysis on the effect of sample size. Section 5 elaborates on this analysis. Palley and Soll (2019) run a second study where they use the same tasks as in Study 1. However they vary subjects' incentives and the sample sizes are much smaller. Thus, their second study will not be considered here.

The second source of data involves the two experimental studies by Wilkening et al. (2021). The first Wilkening et al. (2021) replicates the experiment initially conducted by Prelec et al. (2017). For each U.S. state, subjects are asked if the most populated city is the capital of that state. Prelec et al. (2017) required subjects to pick true or false and report the percentage of other subjects who would agree with them. Wilkening et al. (2021) asked subjects to report probabilistic predictions and meta-predictions on the statement (largest city being the capital city), which allows us to implement the SO algorithm. There are 89 subjects in total and each subject answered 50 questions (one per state). In the second experiment, subjects are presented with U.S. grade school level true/false general science statements. Examples include 'Water boils at 100 degrees Celsius at sea level', 'Materials that let electricity pass through them easily are called insulators' and 'Voluntary muscles are controlled by the cerebrum' etc. The experiments elicits judgments on 500 statements in total. Each subject reports a prediction and a meta-prediction on the probability of a statement being true for 100 such statements. The number of subjects reporting on a given statement varies between 89 to 95. The State Capital and General Knowledge data sets allow us to test the SO algorithm in relatively more practical settings where only binary outcomes can be observed.

4.2 Benchmarks

The benchmarks in testing the SO algorithm can be categorized in two groups. I will first consider *simple benchmarks*, namely the simple average and median prediction. Simple averaging is an easy and intuitive aggregation method. The median forecast is also popular because it is more robust to outliers. These simple aggregation methods do not require meta-predictions, which makes them easier to implement. However, as shown in Section 2 with simple averaging, these methods may produce an inaccurate aggregate judgment. As discussed in Section 1, there exists a growing literature which provides more sophisticated solutions to the aggregation problem utilizing meta-beliefs. I consider three *advanced benchmarks*: Palley and Soll (2019)'s pivoting, Palley and Satopää (2020)'s knowledge-weighting (KW), and Martinie et al. (2020)'s meta-probability weighting (MPW).

The pivoting method first computes simple average of predictions and meta-predictions, $\bar{x}$ and $\bar{z}$ in our notation respectively. Then the mechanism pivots from $\bar{x}$ in different directions. The pivot in the direction of $\bar{z}$ provides an estimate for the shared information while the step in the opposite direction gives an estimate for the average of private signals. These estimates are combined using Bayesian weights to produce the optimal aggregate estimate. The canonical pivoting method requires knowledge of the Bayesian weight ω to determine the optimal pivot size and optimal aggregation. Palley and Soll (2019) propose minimal pivoting (MP) as a simple variant which adjusts $\bar{x}$ by $\bar{x} - \bar{z}$. The adjustment moves the aggregate estimate away from the shared information and alleviates the shared-information problem. MP does not require the knowledge of ω but it only partially corrects for the inconsistency in $\bar{x}$.

The KW mechanism proposes a weighted crowd average as the aggregate prediction. The weights are estimated by minimizing the peer prediction gap, which measures the accuracy of weighted crowds' aggregate meta-prediction in estimating the average prediction. In a similar framework to Section 2, Palley and Satopää (2020) show that minimizing the peer prediction gap is a proxy for minimizing the mean squared error of a weighted aggregate prediction. Intuitively, KW builds on the idea that a weighted crowd that is accurate in predicting others could be more accurate in predicting the unknown quantity itself as well. The mechanism's aggregate prediction is simply the weighted average prediction of such a crowd.

The MPW algorithm aims to construct a weighted average of probabilistic predictions. Martinie et al. (2020) consider a slightly different Bayesian setup where agents receive a private signal from one of the two signal technologies, one for experts and the other for novices. The absolute difference between an agent's optimal prediction and meta-prediction is higher if the agent's signal is more informative. Based on this result, the MPW algorithm assigns weights proportional to the absolute differences between their prediction and meta-prediction. It is expected that agents with more informative private signals receive higher weights and the weighted average becomes more accurate as a result.

Similar to the advanced benchmarks listed above, the SO algorithm relies on an augmented elicitation procedure that elicits meta-predictions in addition to predictions. In contrast, the mechanisms in simple benchmarks do not require information from meta-predictions. Thus we may expect the SO algorithm to significantly outperform simple benchmarks. The advanced benchmarks have similar information demands to the SO algorithm. Thus, comparisons with the advanced benchmarks may produce different results in different tasks.

4.3 Preliminary evidence on overshoot surprises

Section 3 established a relationship between the size and direction of overshoot surprises and prediction errors. The more p_z differs from p_x, the higher the overshoot surprise, suggesting a more inaccurate average prediction. Presence of an overshoot surprise relates to the performance of the SO algorithm as well. Furthermore, we may expect a larger error reduction from using the SO algorithm when $|p_z - p_x|$ is larger.

The Coin Flips data set presents an opportunity to investigate whether overshoot surprises correlate with the inconsistency in the average prediction. In this experiment, both the shared signal s and the unknown probability θ in each coin are generated by the experimenter. Recall that a positive (negative) overshoot surprise is associated with $s > \theta$ ($s < \theta$). We expect no overshoot surprise if $s = \theta$, resulting in $\bar{x}$ being perfectly accurate. Since the information on s and θ is available, we can investigate if this pattern is observed in the sample data. Figure 4 shows the relationship between overshoot surprises and $s - \theta$. Each dot represents an item and the blue lines shows the best linear fit.

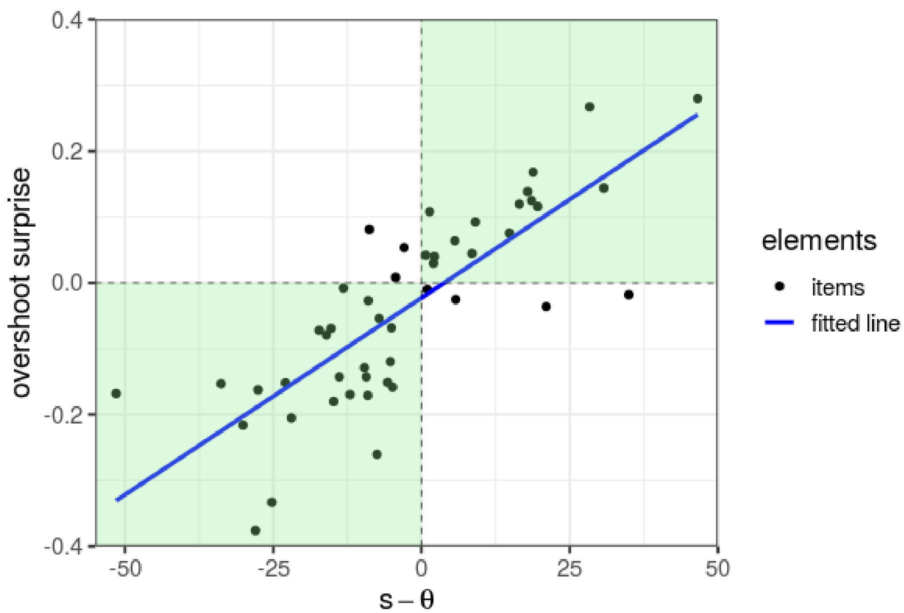


Figure 4: The relationship between $s - \theta$ and overshoot surprises in prediction tasks. Shaded areas show the regions where the sign of $s - \theta$ and overshoot surprise is as predicted by the theory.

The figure shows a strong linear association between $s - \theta$ and sample overshoot surprises. Also observe that most of the points are within the shaded regions. A positive (negative) overshoot surprise is much more likely to occur when $s > \theta$ ($s < \theta$). In addition, the magnitude of overshoot surprise is higher when the difference between s and θ is higher. An overshoot surprise is indeed a strong indicator of the size and direction of the inconsistency in the average prediction. The SO estimator can be thought of as $\bar{x}_N$ adjusted away from the direction of error in $\bar{x}$ where the adjustment is determined by the sign and magnitude of the overshoot surprise. Thus, Figure 4 suggests a potential error reduction from using the SO algorithm. Section 5 explores whether the SO algorithm improves over various benchmarks.

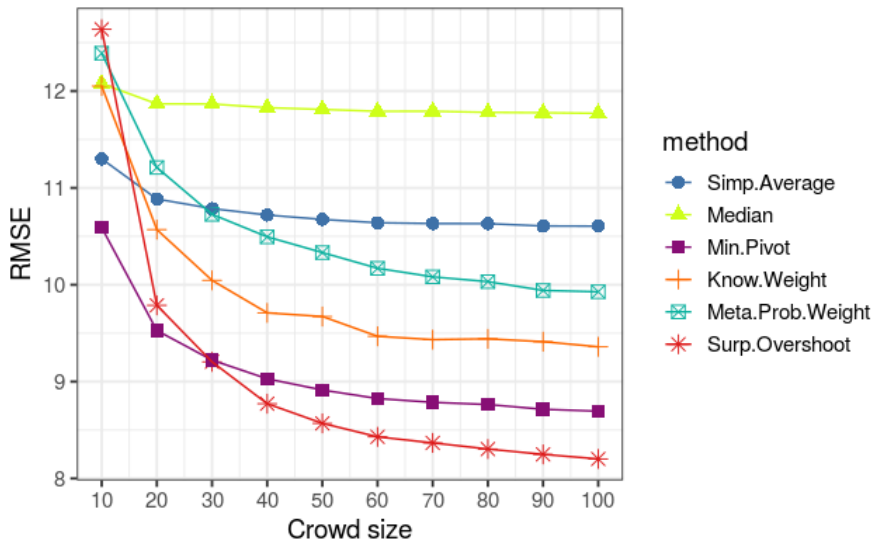
5 Results

5.1 Coin Flips data

In testing the SO estimator with Coin Flips data, I follow a bootstrap approach similar to Palley and Satopää (2020). For each item in a given data set, a subset of subjects of size M is randomly selected to construct a bootstrap sample. For each sample and item, I compute the absolute errors of aggregate predictions from the benchmarks and the SO algorithm. The average of absolute errors across the items gives a measure of the corresponding method's error in that task. This procedure is run 1000 times for each crowd size to obtain 1000 data points on each method's errors for a given crowd size. The observations from bootstrap samples allow us to test for differences in errors between the SO algorithm and a benchmark. I consider two measures for comparison. First, I calculate average RMSE across all iterations for each method. Second, I log transform the errors and calculate pairwise differences for each iteration to construct 95% bootstrap confidence intervals for statistical inference. The differences in log-transformed errors can be interpreted as percentage error reduction. The bootstrap approach also allows us to see the effect of crowd size on the relative performance of the SO algorithm.

Figure 5 presents the results of the bootstrap analysis. Figure 5a depicts the average RMSE across iterations. Figure 5b shows the bootstrap confidence intervals for reduction in log absolute error (the SO estimator vs benchmark). The two panels in 5b depict the error reductions compared to simple and advanced benchmarks respectively.

(a) Average RMSE (across iterations) vs crowd size



(b) Reduction in log absolute error (averaged across items) in Bootstrap samples

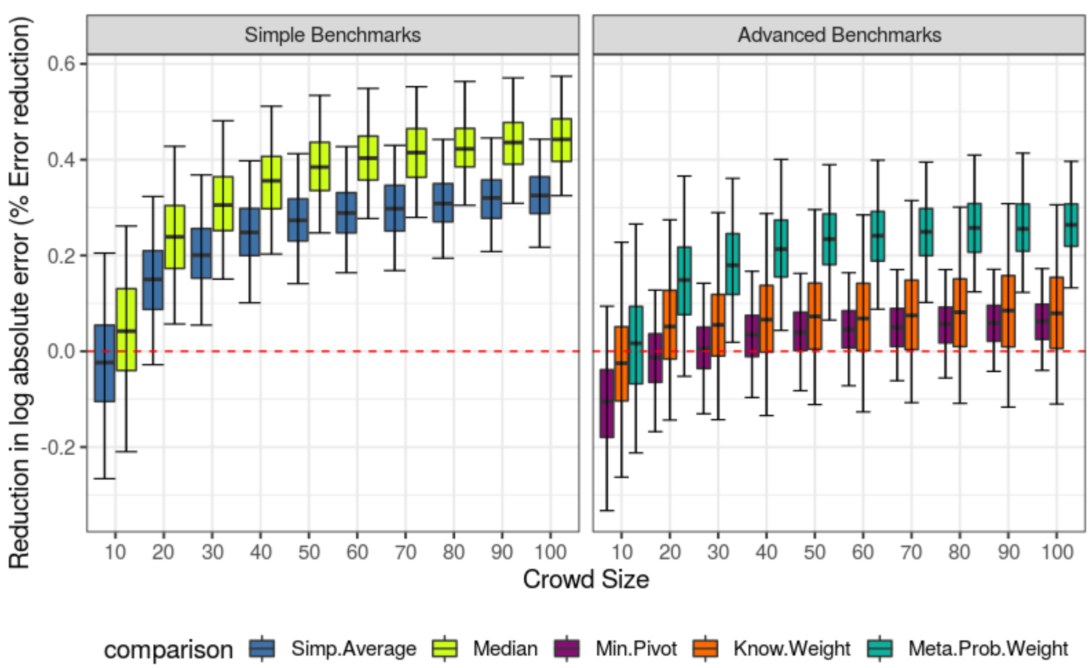


Figure 5: Results of bootstrap analysis on Coin Flips data

Figure 5a shows that the SO algorithm achieves the lowest errors in samples of more than 30 subjects. Observe that increasing the sample size has a stronger effect on the SO estimator. Almost all aggregation methods benefit from larger samples due to the wisdom of crowds effect. For the SO algorithm, benefits of a larger crowd are twofold. Not only the wisdom of crowds effect becomes more pronounced, but also a larger sample of predictions is typically more dispersed and allows a more accurate approximation of the population density of predictions.

Figure 5b indicates that the SO algorithm outperforms the simple benchmarks. We also see that the SO algorithm achieves lower errors in most bootstrap samples than the advanced benchmarks. The differences are significant only for the MPW method. Appendix C provides the 95% bootstrap confidence intervals depicted in Figure 5b. The SO algorithm improves the accuracy by 30-50% relative to the simple benchmarks. The median percentage error reduction with respect to MPW is around 25% in large samples.

The Coin Flips study elicits judgments on a controlled prediction task where all private signals are equally informative and the data generation process of signals induces a distribution of predictions. The following section presents evidence from General Knowledge and State Capital data. In those tasks, subjects may differ in knowledge and the distribution of predictions varies across items.

5.2 State Capital and General Knowledge data

I now consider the State Capital and General Knowledge data sets from the experimental studies in Wilkening et al. (2021). As discussed in Section 4.1, items in these data sets have a binary truth. The analysis in this section tests if the SO algorithm produces a probabilistic estimate closer to the true answer. I follow a similar approach to Budescu and Chen (2015) and Martinie et al. (2020) and calculate transformed Brier scores associated with the aggregate estimates of each method in each data set. The transformed Brier score of a method i in a given data set is defined as

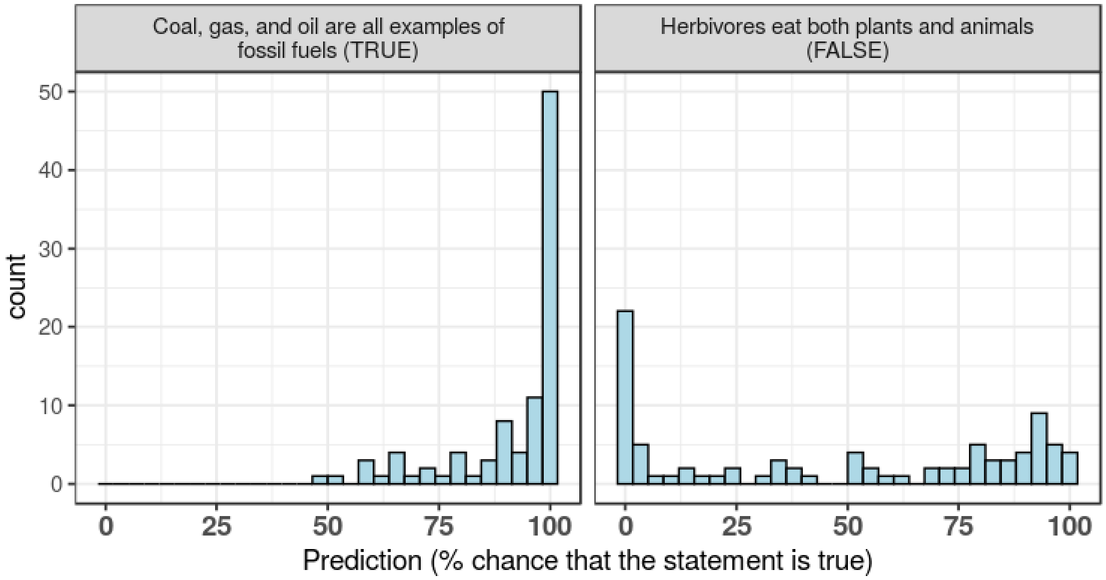
$$S_i = 100 - 100 \sum_{j=1}^{J} \frac{(o_j - x_j^i)^2}{J}$$

where $o_j \in \{0, 1\}$ be the outcome of event j, J is the total number of events in the data set and $x_j^i \in [0, 1]$ is the aggregate probabilistic prediction of method i on event j. The transformed Brier score is strictly proper and assigns a score within $[0, 100]$.

The main question of interest is whether the SO algorithm achieves higher transformed Brier scores than the benchmarks. For statistical inference, I will construct 95% confidence intervals for paired differences in Transformed Brier scores using a bootstrap approach. I generate 1000 bootstrap samples for each method in each data set. A bootstrap sample consists of items sampled with replacement and provides a measurement on S_i. Then, I calculate the paired differences between the scores of the SO estimator and the other benchmarks for each bootstrap sample. Similar to the analysis in Coin Flips data, the 2.5% and 97.5% quantiles of pairwise differences provide a 95% bootstrap confidence interval for testing SO algorithm against a benchmark.

Section 3 discussed why the SO algorithm becomes more effective when predictions are more dispersed. The tasks in General Knowledge and State Capital data sets differ in terms of difficulty and the presence of a strong consensus among the predictions. This allows us to investigate how the extent of disagreement in predictions relates to the relative performance of SO algorithm. To illustrate, consider the two example items from the General Knowledge data in Figure 6 below:

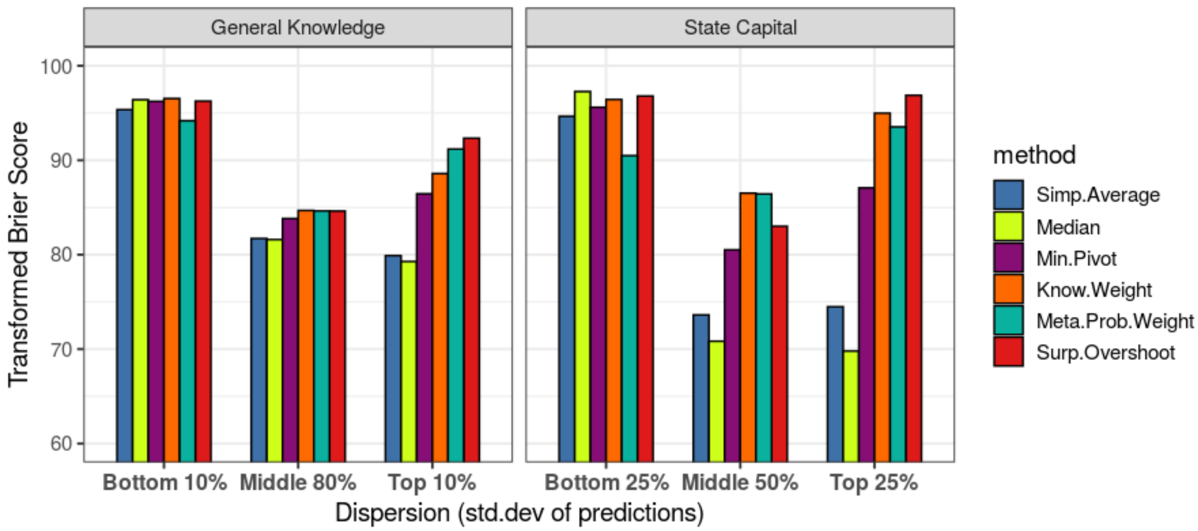
Figure 6: Predictions on two example items from the General Knowledge data



For the item in the left panel, a large proportion of predictions are at 100% and almost all predictions are 50% or higher. The dispersion of predictions is relatively smaller than the item in the right panel, where prediction vary from 0% to 100%. Similar examples can be found in the State Capital data. Given such variety, we can categorize the items in terms of the dispersion of predictions and calculate transformed Brier scores for each category of items. For the main results below, I use standard deviation of predictions as the measure of dispersion in an item. Appendix D replicates the same analysis using kurtosis as the measure and finds very similar results. In the General Knowledge data, I categorize the items in three groups in terms of the standard deviation of predictions: bottom 10%, middle 80% and top 10%. The bottom and top 10% items represent the low and high dispersion items respectively. The State Capital data includes a lower number of items. In order to have a reasonable number of items in each category, the thresholds are set at 25% and 75%. Thus, the categories in the State capital data are bottom 25%, middle 50% and top 25%. The transformed brier score will be calculated separately for each dispersion category.

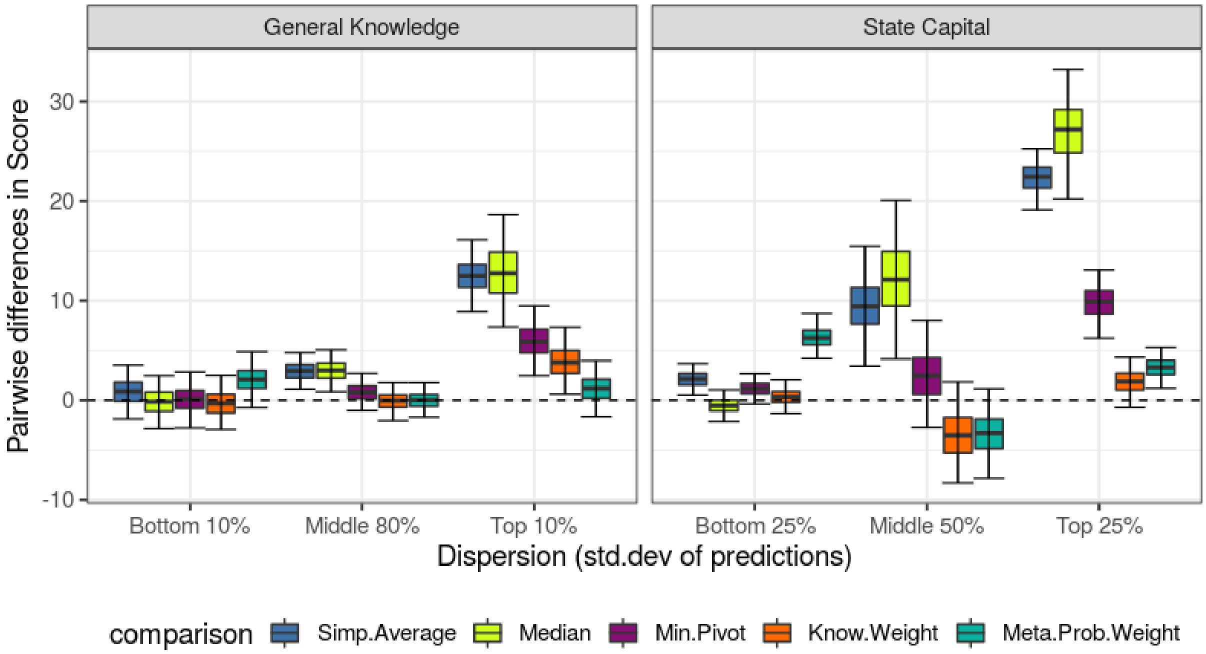
The discussion in Section 3 suggests that the SO algorithm would be relatively more effective in the high-dispersion items. Figure 7 below shows the transformed Brier scores for each method across various levels of dispersion in the General Knowledge and State Capital items.

Figure 7: Transformed Brier Scores in the two data sets



In both data sets, the transformed Brier scores are comparable across all methods in low-dispersion items. Individual predictions are more spread in high-dispersion items, implying a higher frequency of inaccurate judgments. As in Figure 6 (the right panel), there could be many predictions putting a high probability on 'True' even though the correct answer is 'False', reducing the accuracy of simple aggregates. Figure 7 demonstrates that the advanced aggregation algorithms strongly outperform simple aggregates in such cases. In addition, the SO algorithm achieves the best transformed Brier score among items with highest dispersion in either data set. So, in high-dispersion items the SO algorithm is more effective relative to other advanced benchmarks as well. Figure 8 below presents 95% bootstrap confidence intervals for pairwise differences in transformed Brier scores. An observation above the 0-line indicates that the SO estimator achieved a higher transformed Brier score than the corresponding benchmark in that particular bootstrap sample.

Figure 8: Bootstrap differences in Transformed Brier Scores



Appendix C provides the Bootstrap confidence intervals depicted in Figure 8. The confidence intervals show that the SO estimator outperforms simple benchmarks in moderate and high dispersion items. Furthermore, the SO algorithm compares favorably to the advanced benchmarks in high-dispersion items. Similar to the results in the Coin Flips study, the SO algorithm is never worse than the advanced benchmarks while being more accurate in some cases.

To summarize, results indicate that the SO algorithm is relatively more effective in samples of more than 30 experts and when individual predictions disagree greatly, resulting in a more dispersed empirical density of predictions. Section 6 provides a further discussion on the strengths and limitations of the SO algorithm.

6 When and why is the SO algorithm effective?

The findings in Section 5 not only document the effectiveness of the SO algorithm but also provides a 'user's manual' for a DM who intends to use an aggregation algorithm to combine probabilistic judgments. Two important features of the aggregation problem allow a characterization of the situations where the SO algorithm is an effective solution. First, the decision maker knows the size of the sample of experts, which is relevant for the SO algorithm. Second, the empirical density of predictions, which is an input of the SO algorithm, is observable to the DM prior to the resolution of the uncertain event.

Section 5.1 showed that, compared to the benchmarks, the forecast errors of the SO algorithm decrease even more rapidly as the sample size increases. The SO algorithm has the highest average error in samples of 10 experts while in samples of 30 or more experts, the SO algorithm achieves the lowest errors. Intuitively, the SO algorithm is more sensitive to the sample size because it relies on the sample density of predictions. The sample quantiles may overlap in very small samples. As the sample size increases, the sample density becomes more representative of the underlying population density and the quantiles could become more distinct. Then the SO algorithm can produce a more fine-tuned aggregate prediction. The DM will prefer the SO algorithm over the benchmarks if a moderate to large sample of experts is available. In small samples, simple aggregation methods or the MP method may be preferred.

The disagreement between experts is also a factor in the effectiveness of the SO algorithm. Consider a situation where there is a strong consensus among experts: individual predictions are clustered around a certain value (low dispersion). We can imagine two scenarios in which the DM would observe such a pattern. Experts could be highly accurate individually, in which case a simple average of predictions would perform sufficiently well. The lowest categories of dispersion depicted in Figure 7 represent such cases. The simple aggregation methods are as accurate as the more sophisticated aggregation algorithms. In the second scenario, experts predictions are clustered around an inaccurate value. Previous work developed aggregation algorithms to pick a correct answer to a multiple choice question when the majority vote is inaccurate (Prelec et al., 2017; Wilkening et al., 2021). An analogous solution in aggregating probabilistic judgments when most experts are inaccurate may identify a contrarian but accurate prediction and discard other predictions. As discussed in Section 4.2, the KW and MPW mechanisms determine individual weights for aggregation. Other than some extreme cases, these mechanisms are highly unlikely to attach 0 weight to a very high proportion of predictions. The MP method makes an adjustment based on average prediction and meta-prediction. It does not attempt to locate more accurate experts. In theory, the SO algorithm can pick the sample quantile that corresponds to the contrarian prediction. Thus, the DM may prefer the SO algorithm. However, the sample quantiles are close to each other when predictions are highly clustered. Thus, the SO algorithm's adjustment may not be sufficiently extreme. Alternatively, if the DM expects a strong consensus with reasonably well-calibrated individual expert predictions, eliciting the predictions only and using a simple aggregation method could be preferable.

Now consider a situation of high dispersion in predictions instead. Experts disagree in their predictions and some experts are less accurate (ex-post) compared to others. The top dispersion categories in General Knowledge and State Capital studies represent this case. Figure 8 suggests that the SO algorithm not only outperforms the simple aggregation methods, but also it is relatively more effective compared to the advanced benchmarks. The SO algorithm performs well under more disagreement because, similar to the intuition earlier, the sample quantiles become more distinct and the SO algorithm has a higher room for improvement. High dispersion in predictions allows more precision in the SO estimator. Thus, a DM who observes strong disagreement among individual predictions may prefer the SO algorithm when it comes to constructing an aggregate prediction. Note that the aggregation problem is more tricky when experts disagree. The absence of a clear consensus is not uncommon despite the fact that experts typically have shared information through their field of expertise. The SO algorithm is particularly effective in problems where the DM might need an effective aggregation algorithm the most.

The SO algorithm differs from the other alternative aggregation algorithms in its use of the empirical density of predictions. For a given level of overshoot surprise, the absolute difference between the SO estimator and the average prediction depends on the dispersion in the empirical density of predictions. However, the SO algorithm always produces an aggregate estimate that lies within the range of individual predictions. Recall that the MP method uses a fixed step size to adjust the average prediction. In contrast, the SO algorithm's adjustment on the aggregate prediction is informed and restrained by the empirical density. This makes the SO estimator more robust to potential over-adjustments, which would reduce the accuracy even when the aggregate estimate is adjusted in the correct direction.

7 Conclusion

Decision makers frequently face the problem of predicting the likelihood of an uncertain event as part of a decision making process. Leveraging the collective wisdom of many experts has been shown to be a promising solution. However, the use of collective wisdom is not a trivial solution because there are typically no general guidelines on how individual judgments should be aggregated for maximum accuracy. Experts typically have shared information through their training, past realizations, knowledge of the same academic works, etc. In such cases, the simple average of predictions exhibits the shared-information problem (Palley and Soll, 2019). Recent work developed aggregation algorithms that rely on an augmented elicitation procedure (Prelec, 2004; Prelec et al., 2017; Palley and Soll, 2019; Palley and Satopää, 2020; Wilkening et al., 2021). These algorithms use individuals' meta-beliefs to aggregate predictions more effectively. This paper follows a similar approach and proposes a novel algorithm to aggregate probabilistic judgments on the likelihood of an event. The Surprising Overshoot algorithm uses experts' probabilistic meta-predictions to aggregate their probabilistic predictions. The SO algorithm utilizes the information in meta-predictions and the empirical density of predictions to produce a consistent estimator of the unknown probability of an event.

Experimental evidence shows that the SO algorithm consistently outperforms simple averaging and median prediction. Furthermore, the SO algorithm compares favorably to alternative aggregation algorithms that elicit meta-beliefs (Palley and Soll, 2019; Palley and Satopää, 2020; Martinie et al., 2020). The SO algorithm is particularly effective when the empirical density of predictions is highly dispersed. Such high dispersion is more likely in moderate to large samples of experts and in prediction tasks where experts disagree in their individual assessment.

In practice, a DM is more likely to need a judgment aggregation algorithm when expert predictions lack a clear consensus. Experts frequently disagree in their judgments even though they have overlapping information. In such decision problems, the DM finds herself with conflicting expert judgments with no straightforward way to combine them. Because of its effectiveness in aggregating disagreeing judgments, the SO algorithm is especially powerful in such challenging aggregation problems. The dispersion in predictions that result from the disagreement among experts works in the algorithm's favor.

References

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

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (2008). *A first course in order statistics*. SIAM.

Budescu, D. V. and Chen, E. (2015). Identifying Expertise to Extract the Wisdom of Crowds. *Management Science*, 61(2):267–280.

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.

Clemen, R. T. and Winkler, R. L. (1986). Combining economic forecasts. *Journal of Business & Economic Statistics*, 4(1):39–46.

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.

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

Kahneman, D., Slovic, S. P., Slovic, P., and Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological review*, 80(4):237.

Larrick, R. P. and Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science*, 52(1):111–127.

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. In Krueger, J. I., editor, *Frontiers of social psychology. Social judgment and decision making*, pages 227–242. Psychology Press.

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.

Palan, S., Huber, J., and Senninger, L. (2019). Aggregation mechanisms for crowd predictions. *Experimental economics*, pages 1–27.

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.

Prelec, D. (2004). A Bayesian Truth Serum for Subjective Data. *Science*, 306(5695):462–466.

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

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. Doubleday & Co, New York, NY, US.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.

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*.

Winkler, R. L., Grushka-Cockayne, Y., Lichtendahl, K. C., and Jose, V. R. R. (2019). Probability forecasts and their combination: A research perspective. *Decision Analysis*, 16(4):239–260.

Appendices

A Proofs

A.1 Theorem 1

Let agent $i \in \{1, 2, \ldots, N\}$ be an arbitrary agent. Suppose all agents $j \in \{1, 2, \ldots, N\} \setminus \{i\}$ report truthfully, i.e. $(x_j, z_j) = (E[\theta|s, t_j], E[\bar{x}_{-j}|s, t_j]$ where $\bar{x}_{-j}$ represents the average prediction of all agents excluding j. Truthful reporting is a Bayesian Nash equilibrium if $(x_i, z_i) = (E[\theta|s, t_i], E[\bar{x}_{-i}|s, t_i]$ is agent i's best response.

Let $(x_i^*, z_i^*) = \arg\max_{x_i, z_i} E[\pi_i|s, t_i]$ denote the optimal prediction and meta-prediction that maximizes agent i's expected score given $\{s, t_i\}$ and truthful reporting from other agents. Note that $E[\pi_i|s, t_i] = E[\pi_{xi}|s, t_i] + E[\pi_{zi}|s, t_i]$. Agent i's prediction does not affect $E[\pi_{zi}|s, t_i]$ as it is completely determined by z_i and $\bar{x}_{-i}$. Similarly, $E[\pi_{xi}|s, t_i]$ is determined by x_i and the realization of Y. Thus agent i's meta-prediction has no effect on $E[\pi_{xi}|s, t_i]$. Thus, agent i's maximization problem is separable where $x_i^* = \arg\max_{x_i} E[\pi_{xi}|s, t_i]$ and $z_i^* = \arg\max_{x_i} E[\pi_{zi}|s, t_i]$. Recall that π_{xi} and π_{zi} are maximized at θ and $\bar{x}_{-i}$ respectively. Then, $x_i^* = E[\theta|s, t_i]$ and $z_i^* = E[\bar{x}_{-i}|s, t_i]$. Truthful report $(x_i, z_i) = (E[\theta|s, t_i], E[\bar{x}_{-i}|s, t_i]$ is agent i's best response, which completes the proof.

A.2 Theorem 2

The SO estimator x^{SO} is consistent if x^{SO} is an asymptotically unbiased estimator of θ and $Var(x_N^{SO}) \to 0$ as $N \to \infty$. The proof will establish asymptotic unbiasedness and zero asymptotic variance.

Asymptotic unbiasedness. Let $B(x_N^{SO}) = \lim_{N\to\infty} E[x_N^{SO} - \theta]$ be the asymptotic bias of the SO estimator x_N^{SO}. From Lemma 1, we can substitute for θ and write $B(x_N^{SO}) = \lim_{N\to\infty} E[x_N^{SO} - F^{-1}(1 - p_z)] = 0$. Then, x^{SO} is asymptotically unbiased if we establish that $B(x_N^{SO}) = 0$. For any given $q \in [0, 1]$, let $\hat{x}_N$ be the qth sample quantile where $F^{-1}(q)$ is the qth population quantile. We have the following result (Theorem 8.5.1 in Arnold et al.

(2008)):

$$\sqrt{N} f(F^{-1}(q)) \frac{\hat{x}_N - F^{-1}(q)}{\sqrt{q(1-q)}} \xrightarrow{d} N(0,1) \tag{6}$$

Thus, $\hat{x}_N$ is an asymptotically unbiased estimator of $F^{-1}(q)$. Recall that x_N^{SO} is the $1 - \hat{p}_{zN}$ sample quantile. Then, x_N^{SO} is an asymptotically unbiased estimator of the population quantile at $1-\hat{p}_{zN}$, given by $F^{-1}(1-\hat{p}_{zN})$. Then, we have $\lim_{N\to\infty} E[x_N^{SO} - F^{-1}(1-\hat{p}_{zN})] = 0$. We can rewrite $B(x_N^{SO})$ as follows:

$$\begin{aligned} B(x_N^{SO}) &= \lim_{N\to\infty} E[x_N^{SO} - F^{-1}(1-p_z)] \\ &= \lim_{N\to\infty} E[x_N^{SO} - F^{-1}(1-\hat{p}_{zN}) + F^{-1}(1-\hat{p}_{zN}) - F^{-1}(1-p_z)] \\ &= \lim_{N\to\infty} E[x_N^{SO} - F^{-1}(1-\hat{p}_{zN})] + \lim_{N\to\infty} E[F^{-1}(1-\hat{p}_{zN}) - F^{-1}(1-p_z)] \end{aligned} \tag{7}$$

Since $\lim_{N\to\infty} E[x_N^{SO} - F^{-1}(1-\hat{p}_{zN})] = 0$, the first additive term in Equation 7 becomes zero. Also recall that $\hat{p}_{zN} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(z_i > \bar{x}_N)$ and p_z represents the population proportion of agents i such that $z_i > \bar{x}$. For any given agent i, the occurrence of $z_i > \bar{x}_N$ can be considered as a Bernoulli process with probability p_z. Then, the sample instances of meta-predictions that overshoot the average prediction follows a Binomial process where $N\hat{p}_{zN} \sim Binom(N, p_z)$, which implies $\hat{p}_{zN} \xrightarrow{p} p_z$. Since F is a continuous function, we also get $F^{-1}(1-\hat{p}_{zN}) \xrightarrow{p} F^{-1}(1-p_z)$ from the continuous mapping theorem, implying that $\lim_{N\to\infty} E[F^{-1}(1-\hat{p}_{zN}) - F^{-1}(1-p_z)] = 0$. Therefore, we get $B(x_N^{SO}) = 0$.

Zero asymptotic variance. As discussed above, Equation 6 applies to the SO estimator. Then, $Var(x_N^{SO})$ is in the order of $\sqrt{N}$, implying that the asymptotic variance satisfies $\lim_{N\to\infty} Var(x_N^{SO}) \to 0$.

Since x_N^{SO} satisfies asymptotic unbiasedness and zero asymptotic variance, it follows that x_N^{SO} is a consistent estimator of θ.

B Mixed sample of experts and non-experts

Without loss of generality, let agents $i \in \{1, 2, \ldots, K\}$ be the *experts* who observe both the shared signal and a private signal. Agents $i \in \{K+1, K+2, \ldots, N\}$ are *non-experts* observe the shared signal s only. Then,

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

Also, we have $z_i = (1-\omega)s + \omega x_i$ for $i \in \{1, 2, \ldots, K\}$ while $z_i = s$ for others. The crowd average is given by $\bar{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_i = (1-\omega)s + \omega\frac{1}{N}\sum_{i=1}^{N} t_i$.

Note that for the experts, equations 4 and 5 hold. Consider $i > K$. We have $x_i > \bar{x}_N$ iff $s > \frac{1}{N}\sum_{i=1}^{N}$. Then, in the limit $x_i > \bar{x}$ iff $s > \theta$. Since $z_i = x_i = s$, we also have $z_i > \bar{x}$ iff $x_i = s > \theta$. So, prediction and meta-prediction of an agent i who observed s only jointly overshoot or undershoot $\bar{x}$. Nevertheless, $z_i > \bar{x} \iff x_i > \theta$ holds, implying that if we define p_x and p_z the same way they are defined in the main text, Lemma 1 applies. Then, the Theorem applies as well.

C Bootstrap confidence intervals

Table C1: 95% Confidence intervals depicted in Figure 5b

C.Size	Comparison	Low.B.	Upp.B.	C.Size	Comparison	Low.B.	Upp.B.
10	Simp.Average	-0.27	0.20	60	Simp.Average	0.16	0.43
10	Median	-0.21	0.26	60	Median	0.28	0.55
10	Min.Pivot	-0.33	0.09	60	Min.Pivot	-0.07	0.16
10	Know.Weight	-0.26	0.23	60	Know.Weight	-0.13	0.28
10	Meta.Prob.Weight	-0.21	0.27	60	Meta.Prob.Weight	0.09	0.40
20	Simp.Average	-0.03	0.32	70	Simp.Average	0.17	0.43
20	Median	0.06	0.43	70	Median	0.28	0.55
20	Min.Pivot	-0.17	0.13	70	Min.Pivot	-0.06	0.17
20	Know.Weight	-0.14	0.27	70	Know.Weight	-0.11	0.31
20	Meta.Prob.Weight	-0.05	0.37	70	Meta.Prob.Weight	0.10	0.39
30	Simp.Average	0.05	0.37	80	Simp.Average	0.19	0.44
30	Median	0.15	0.48	80	Median	0.30	0.56
30	Min.Pivot	-0.13	0.14	80	Min.Pivot	-0.06	0.17
30	Know.Weight	-0.14	0.29	80	Know.Weight	-0.11	0.30
30	Meta.Prob.Weight	0.02	0.36	80	Meta.Prob.Weight	0.12	0.41
40	Simp.Average	0.10	0.40	90	Simp.Average	0.21	0.45
40	Median	0.20	0.51	90	Median	0.31	0.57
40	Min.Pivot	-0.10	0.17	90	Min.Pivot	-0.04	0.17
40	Know.Weight	-0.13	0.29	90	Know.Weight	-0.12	0.31
40	Meta.Prob.Weight	0.04	0.40	90	Meta.Prob.Weight	0.12	0.41
50	Simp.Average	0.14	0.41	100	Simp.Average	0.22	0.44
50	Median	0.25	0.53	100	Median	0.32	0.57
50	Min.Pivot	-0.08	0.16	100	Min.Pivot	-0.04	0.17
50	Know.Weight	-0.11	0.30	100	Know.Weight	-0.11	0.31
50	Meta.Prob.Weight	0.07	0.39	100	Meta.Prob.Weight	0.13	0.40

Table C2: 95% Bootstrap confidence intervals depicted in Figure 8, General Knowledge data

Comparison	Dispersion	Low.B.	Upp.B.
Simp.Average	Bottom 10%	-1.87	3.54
Median	Bottom 10%	-2.84	2.46
Min.Pivot	Bottom 10%	-2.78	2.85
Know.Weight	Bottom 10%	-2.94	2.49
Meta.Prob.Weight	Bottom 10%	-0.75	4.88
Simp.Average	Middle 80%	1.13	4.78
Median	Middle 80%	0.87	5.07
Min.Pivot	Middle 80%	-1.04	2.71
Know.Weight	Middle 80%	-2.03	1.76
Meta.Prob.Weight	Middle 80%	-1.69	1.76
Simp.Average	Top 10%	8.93	16.13
Median	Top 10%	7.35	18.66
Min.Pivot	Top 10%	2.46	9.47
Know.Weight	Top 10%	0.62	7.33
Meta.Prob.Weight	Top 10%	-1.63	3.98

Table C3: 95% Bootstrap confidence intervals depicted in Figure 8, State Capital data

Comparison	Dispersion	Low.B.	Upp.B.
Simp.Average	Bottom 25%	0.53	3.69
Median	Bottom 25%	-2.13	1.06
Min.Pivot	Bottom 25%	-0.37	2.67
Know.Weight	Bottom 25%	-1.32	2.05
Meta.Prob.Weight	Bottom 25%	4.20	8.74
Simp.Average	Middle 50%	3.44	15.48
Median	Middle 50%	4.15	20.07
Min.Pivot	Middle 50%	-2.74	8.02
Know.Weight	Middle 50%	-8.31	1.83
Meta.Prob.Weight	Middle 50%	-7.85	1.17
Simp.Average	Top 25%	19.16	25.25
Median	Top 25%	20.22	33.22
Min.Pivot	Top 25%	6.26	13.09
Know.Weight	Top 25%	-0.72	4.32
Meta.Prob.Weight	Top 25%	1.23	5.31

D Additional figures on dispersion

Figure D1: Transformed Brier Scores (measure of dispersion: kurtosis)

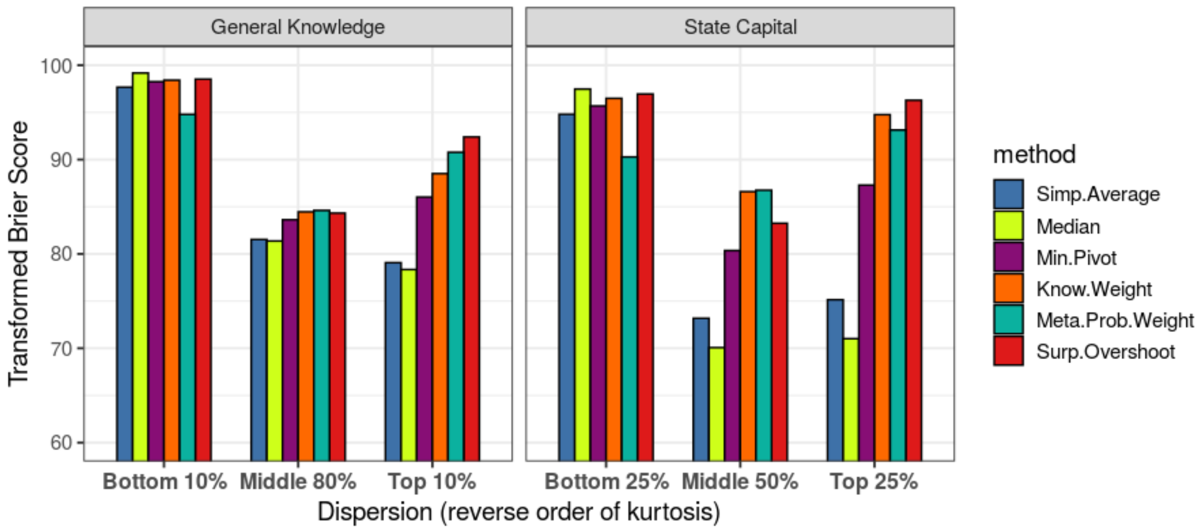


Figure D2: Bootstrap differences in Transformed Brier Scores (measure of dispersion: kurtosis)