Follow the money, not the majority: A mechanism predicting unresolvable events*

Aurélien Baillon† Benjamin Tereick‡ Tong V. Wang§

January 7, 2025

Abstract

Mechanisms such as scoring rules and prediction markets can be used to incentivize truthful signal reporting and to aggregate signals. However, they are fundamentally limited by the fact that uncertainty should resolve. In this paper, we develop a mechanism, based on Bayesian markets, that incentivizes accuracy and aggregates information for unresolvable events. Participants decide whether to endorse a statement and trade an asset whose value depends on the endorsement rate. The respective payoffs of buyers and sellers indicate whose endorsement to trust. We demonstrate theoretically and illustrate empirically that "following the money" outperforms selecting the majority endorsement.

keywords: signal elicitation; wisdom of crowds; expert aggregation; truth-telling incentives.

*This research was made possible by European Research Council Starting Grant 638408 BayesianMarkets. We are grateful to Drazen Prelec, Aniol Llorente-Saguer, Marie Claire Villeval, and Sander Renes for their comments. Ethical approval was obtained from the internal review board of Erasmus School of Economics for experimental research, under file number ETH2122-0805, on July 7, 2022.

†baillon@em-lyon.com, emlyon business school, CNRS, Université Lumière Lyon 2, Université Jean Monnet Saint-Etienne, GATE, 144 rue Jean Jaurès, 69007 Lyon, France

‡benjamin.tereick@openphilanthropy.org, Open Philanthropy Project.

§t.wang@dufe.edu.cn, Center for Industrial and Business Organization and Institute for Advanced Economic Research, Dongbei University of Finance and Economics.

1. Introduction

Consider a group of researchers conducting a multi-lab or multi-analyst study. They independently test the same hypothesis, either by each conducting an experiment (multi-lab) or analyzing existing data (multi-analyst), and they then report the results. In essence, running an experiment or analyzing data in a specific way can be seen as drawing a binary signal ("support" or "falsify") about the state of nature (whether the stated hypothesis is true or not). In this example, there is no guarantee that the researchers will all draw the same signal. In fact, they are likely to obtain surprisingly diverging, and sometimes contradicting, results (Klein et al., 2018; Menkveld et al., 2023).

In many cases, uncertainty will never resolve. In other words, the state of nature will never be observed, and it cannot be inferred in any other way than by aggregating the researchers' reported signals. The unresolvability of the state of nature creates two problems: the "incentive problem" and the "aggregation problem".

First, how can we incentivize truthful signal reporting when the true state of nature cannot be independently observed, and signals are private? Traditional mechanisms like scoring rules or prediction markets rely on the resolution of uncertainty to determine payoffs. Without an observable state of nature, no payoff can take place.[1] An alternative approach is to define the state of nature as a function of the submitted reports. While this might seem like a solution, it introduces a strategic challenge: participants may focus on aligning their reports with what they believe others will report, rather than truthfully reporting their own private signals. This "beauty contest" effect undermines the objective of eliciting honest, individual information.

Besides the incentive problem, we furthermore face an "aggregation problem": which signals to trust if they diverge? The obvious candidate is the majority signal,

[1] Even more advanced methods controlling for differences in endowment (Jaffray and Karni, 1999) or differences in risk attitudes (Andersen et al., 2014) rely on observable states of nature.

but there is no guarantee that it is the best approach. In some extreme cases, a single falsification among many attempts should lead to the rejection of the hypothesis. This would be the case under a strictly Popperian scientific methodology (Popper, 1959) or when validating a mathematical statement, where a single counterexample would be sufficient to establish its falsity. Obviously, in most scientific endeavors, especially in the social sciences, experiments can be noisy and one may expect some experiments falsifying a hypothesis even if it is true. However, the main argument remains. Some signals that are more difficult to get should drive the conclusion even if they are a minority.

In this paper, we show that a mechanism proposed to address the incentive problem can be modified to simultaneously solve the aggregation problem. Baillon (2017) designed a "Bayesian market" where agents endorse a statement or not after observing a signal, and then trade an asset whose value is determined by the total endorsements. Those who endorse the statement are offered to buy the asset from a *center* at price p, where p is randomly drawn. Essentially, buying the asset is betting that more than $p\%$ of others will endorse the statement. Those not endorsing the statement can sell the asset to the center. Baillon (2017) showed that such a "Bayesian market" provides incentives for agents to endorse the statement consistent with their signal, avoiding both a beauty-contest effect and the no-trade theorem (Milgrom and Stokey, 1982) through the intermediary role of the center.

By modifying this mechanism—making the price individualized, independently drawn for each agent—we show that Bayesian markets have desirable properties with respect to aggregation while keeping the incentive property intact. With sufficiently many agents, those with the signal that indicates the actual state of nature, and only them, will make a profit. Hence, by "following the money", we can infer the state of nature without relying on what the majority reports.

The intuition of our result is based on an argument put forward by Prelec et al. (2017). If signals are correlated with the states of nature, there will be more signals supporting a state of nature when this state is the actual one than when it is not, and

therefore, than we would have expected ex ante. Prelec et al. (2017) proposed the *surprisingly popular algorithm* (SPA) in which agents are asked to endorse a state and predict the rate of endorsement. The algorithm picks the state that is more often endorsed than agents predicted. Prelec et al. (2017) demonstrated theoretically and experimentally that this approach improves upon following the majority and confidence-weighted aggregation.[2]

We combine the ideas behind Bayesian markets and the SPA and show that in equilibrium, "following the money" in a Bayesian market offers the same aggregation benefits as the SPA does under truthful reporting. In contrast to the SPA, however, Bayesian markets provide monetary incentives for accuracy, a crucial feature for high-stakes domains in which information acquisition is costly. Moreover, Bayesian markets require less information than the SPA. Agents only reveal their signal and make a binary trading decision. It is known in the literature that the incentives challenge requires asking more than signals (e.g., Radanovic and Faltings, 2013, Theorem 1) when the state of nature is unobservable. Virtually all alternative methods ask predictions on top of signals (Prelec, 2004; Witkowski and Parkes, 2012; Radanovic and Faltings, 2013, 2014; Cvitanić et al., 2019). Bayesian markets use trades to incentivize truthful revelation of signals. With this little extra information, we can recover agents' predictions by fitting supply and demand curves for the asset, and even infer the whole signal technology, thereby solving the aggregation challenge. In addition to requiring less information than the method of Prelec et al. (2017), our market approach also opens up the possibility of continuous markets, extending prediction markets to unresolvable events.

The next section of the paper introduces the theoretical setting and the mechanism. We analyze payoffs at the equilibrium and show how the endorsement of those with positive payoffs indicates the actual state of nature. If the statement is true, those endorsing it can make a profit from betting on others' endorsement rate. If

[2]See Wilkening et al. (2022), Peker (2022), and Palley and Satopää (2023) for follow-up methods adapted to probabilistic forecasts.

it is not true, those rejecting it can make a profit. The profits realize even in the absence of uncertainty resolution, because bets are based on endorsement rates, not on states.

Section 3 describes a preregistered experiment we ran on a large sample of US students. We used a task developed by Tereick (2020) that ensures that the informational assumptions of the model are satisfied. Under these assumptions, homo economicus would behave exactly as our model predicts. Our experiment allowed us to test whether our method also worked for homo sapiens, without having to worry whether the informational part of the model perfectly described the reality. We compared our method to the majority opinion and to the SPA. Despite using less information than the SPA, our method had comparable accuracy rates. Both methods substantially improved upon majority.

2. Theory

2.1. Setting

Let $\{Y, N\}$ be the *state space*, with Y and N the two possible *states of nature*. For instance, these two states can represent whether a statement is true or not. Which state S we are in is assumed to be unobservable (or equivalently, uncertainty does not resolved).

A group of n expert *agents*, however, has private information about the state. The *common prior* of the agents is that the probability of state Y is r. Each agent gets a *private signal* $s_i \in \{0, 1\}$, with sampling probabilities $P(s_i = 1 \,|\, Y) = \omega_Y$ and $P(s_i = 1 \,|\, N) = \omega_N$. Signals are independent conditionally on the state, i.e. $P(s_i = 1 \,|\, S, s_j) = \omega_S$ for all $S \in \{Y, N\}$ and $j \neq i$.[3] We assume $\omega_Y > \omega_N$. This implies that signals are informative about the state of nature, $s_i = 1$ providing

[3]In other words, signals are independent and identically distributed given the state, but the latter is uncertain. The absence of correlation between signals implies that agents will not exhibit correlation neglect, unlike studied by Enke and Zimmermann (2019).

support for Y and $s_i = 0$ for N. We do not require $\omega_Y > 0.5 > \omega_N$, which would be necessary for the majority of signals to be correct (in an infinite group of agents). The assumption $\omega_Y > \omega_N$ is as mild as can be. Equality would mean that s_i is non informative and therefore, all agents would stick to the prior belief r. The opposite inequality would simply change the interpretation of the signal ($s_i = 0$ providing support for Y and $s_i = 1$ for N). Together, we call the triplet $\langle \omega_Y, \omega_N, r \rangle$ a *signal technology*.

Using Bayesian updating, agents form posterior beliefs about the actual state according to

$$r_1 \equiv P(Y \mid s_i = 1) = \frac{r\omega_Y}{r\omega_Y + (1-r)\omega_N}; \tag{1}$$

$$r_0 \equiv P(Y \mid s_i = 0) = \frac{r(1-\omega_Y)}{r(1-\omega_Y) + (1-r)(1-\omega_N)}. \tag{2}$$

For simplicity, we assume that ω_Y, ω_N, and r are such that $r_1 > 0.5$ and $r_0 < 0.5$. It allows us to equate an agent's signal with the state the agent believes more likely to be the actual state (the state they endorse, if they are honest). If this assumption is not satisfied, signals would be informative but a single signal would not suffice to reverse one's belief. A sufficient condition for this assumption is $r = 0.5$, as used in our experiment. The reason we focus on endorsements rather than signals in this paper is because in many practical applications, it will be much easier for respondents to identify which state of the world they deem more likely, rather than to articulate the source of this belief.[4]

Apart from the agents' posterior beliefs about states, we can also infer posterior

[4]Sometimes, one may however be interested in eliciting beliefs where there is agreement about the most likely state. This may be for instance, when predicting catastrophic events, where a probability of, say 5% instead of 0.01%, makes a huge difference. There are two ways of using Bayesian markets in such situations. First, the market can ask about the signal directly. In this case, the assumption $r_0 < 0.5 < r_1$ can be dropped, and all results of this paper still hold - at the cost of the practical difficulty of asking respondents about their information sources. Alternatively, one can replace endorsements by the question "Do you think state Y is more likely than the average person in our sample thinks?". Again, in this case, it is not needed that $r_0 < 0.5 < r_1$.

expectations about the proportion of agents who received signal 1 in the population. We denote the actual value of this proportion by ω. Since the expectation of a proportion under random sampling equals the sampling probabilities, agents who received signal 1 expect ω to be

$$\overline{\omega}_1 \equiv E\left[\omega \mid s_i = 1\right] = r_1\,\omega_Y + (1 - r_1)\,\omega_N, \tag{3}$$

whereas agents with signal 0 expect

$$\overline{\omega}_0 \equiv E\left[\omega \mid s_i = 0\right] = r_0\,\omega_Y + (1 - r_0)\,\omega_N. \tag{4}$$

A *center* wants to find out which state we are in (the *actual* state). This center can be a policy maker consulting experts, but could just as well be an employer querying employees or a scientific association surveying its members. We make the usual assumption that the signal technology is common knowledge among the agents. However, as in Prelec (2004), Baillon (2017), Prelec et al. (2017), and Cvitanić et al. (2019), the center does not know the signal technology. This setting has two implications. First, the center cannot only ask signals. Observing a proportion of signals 1 would not suffice to infer the actual state. Second, the center cannot only ask the signal technology. Even the agents who know the signal technology and their own signal cannot infer the actual state with certainty.

The problem faced by the center is a mechanism design problem, i.e., creating an institution to recover the state of nature given these information constraints. Expressed in the terms of our model, the incentive and aggregation problem can be stated as follows. Each agent will report an endorsement e_i, where $e_i = 1$ denotes that agent i endorses state Y and $e_i = 0$ that i endorses state N. The center wants to reward the agents in such a way that it becomes profitable for them to endorse a state if and only if they believe it more likely to be the actual one. Furthermore, upon learning the endorsements $e_1, \ldots, e_n$, the center selects one of the two states, and wishes to maximize the probability that it is the actual one. Since the state S is unobservable and the signal technology is unknown to the center, it is not

possible to make the payments or selection of a state dependent on the actual state, nor the selection of the state dependent on the parameters ω_Y and ω_N. Thus, it is impossible to use traditional methods to elicit agents' signals or beliefs because the signals are private (impossible to directly reward truth-telling) and the beliefs are about unresolvable states Y and N (bets and scoring rules cannot be applied). Second, even knowing signals or beliefs would not enable the center to determine the state of nature because the center does not know the values for ω_Y and ω_N. In other words, for anyone unaware of the signal technology, observing 20% of signal 1 does not say which state we are in.

The next subsection introduces the mechanism, called a Bayesian market.

2.2. Bayesian market

The center and each agent i trade an asset whose *settlement value* v is defined as the share of agents endorsing state Y, i.e.,

$$v = \frac{\sum_{j=1}^{n} e_j}{n}.$$

The center organizes a *Bayesian market* for these assets:

1. Agents simultaneously report e_i to the center only.

2. For each agent i, the center draws a price p_i from a uniform distribution over $(0, 1)$ and proposes the following trade to the agent, and the agent can decide to take up the offer ($d_i = 1$) or not ($d_i = 0$):

   (a) If $e_i = 1$, agent i can buy the asset at price p_i from the center;

   (b) If $e_i = 0$, agent i can sell the asset at price p_i to the center.

3. All endorsements e_i and buying/selling decisions d_i are revealed.

4. (a) If an agent decided to buy at price p_i, then the trade occurs if there exists another agent j selling at $p_j \leq p_i$.

(b) If an agent decided to sell at price p_i, then the trade occurs if there exists another agent j buying at $p_j \geq p_i$.

5. Those agents who bought the asset collect v and pay p_i; those who sold it collect p_i and pay v.

Step 2 differs slightly from the mechanism proposed in Baillon (2017) in which a single price p is drawn for all agents. The motivation for the change is to learn as much as possible from the decisions of different agents. When only a single price is drawn and, e.g., all potential buyers reject the trade, the center only learns that the price was larger than the buyers' reservation price, but not by how much. An alternative would be to directly ask agents for their reservation prices. The center could then draw only one random price p for all agents. This would correspond to the Becker-DeGroot-Marshak mechanism (Becker et al., 1964), but with the trading rule (step 4) in place. The advantage of binary decisions in step 2 is that they require less information from the agents, and therefore less cognitive effort. It is easier to buy/sell at a given price (equivalently, to take/reject a bet on the asset value) than to report a reservation price.[5]

Our mechanism as stated induces a game played among the agents. In this game, a *strategy profile* is a collection $(e, d) = ((e_1, d_1), ..., (e_n, d_n))$, where e_i determines which state individual i is going to endorse depending on the signal s_i, and the trading strategy d_i assigns to each possible signal a range of prices in the $(0, 1)$-interval which i is going to accept when receiving a buy or sell offer from the center. Note that this definition of strategies precludes mixed strategies and the existence of an external coordination device among agents, so that the actual endorsements

[5]Asking for reservation prices, however, has advantages regarding the logistical aspects of practical implementation: In our design, a random price must be drawn for every respondent. When asking for reservation prices, respondents can be contacted by a pen and paper survey in which they submit their reservation prices and a public price is later credibly drawn. Whether these practical considerations outweigh the cognitive simplicity of a binary decision, will depend on the application.

made by agents are fully determined by their signal and strategy. In Section 5, we discuss this strategy restriction in light of our empirical results.

The mechanism assigns a payoff $U_i(e, d)$ to each agent. Importantly, these payoffs cannot depend on the actual state of nature S or its ω_S. A *Bayesian Nash equilibrium* of the induced game means that, conditioning on their signal, no agent expects a higher payoff by moving to another strategy, i.e.,

$$E\left[U_i\left(e, d\right) \mid s_i\right] \geq E\left[U_i\left(\left(e_1, d_1\right), ..., \left(e_i', d_i'\right), ..., \left(e_n, d_n\right)\right) \mid s_i\right]$$

for any $(e_i', d_i') \neq (e_i, d_i)$ and all signal realizations $s_i \in \{0, 1\}$. We further say that a strategy profile is *truthful* or, equivalently, that there is *truth-telling*, if $e_i(1) = 1$ and $e_i(0) = 0$ for any agent i.

We assume that all agents are risk-neutral[6] and care only about their own monetary payoff, so that $U_i(e, d)$ is just i's monetary payoff. If $e_i = 1$, agent i is potentially a buyer, and we denote by $\pi_1(v, p_i)$ agent i's monetary payoff if deciding to buy ($d_i = 1$), as a function of the asset value v and individualized buying price p_i. Then

$$\pi_1(v, p_i) = \begin{cases} v - p_i & \text{if trade happens;} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Symmetrically, $\pi_0(v, p_i)$ denotes agent i's monetary payoff as a potential seller if deciding to sell ($d_i = 0$):

$$\pi_0(v, p_i) = \begin{cases} p_i - v & \text{if trade happens;} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

2.3. Equilibrium behavior

From here, we assume that n is infinite. Three simplifications come with an infinite group of expert agents, which together, imply that the asset value is simply ω_Y or

[6]The assumption of risk neutrality is rather common in the literature on expert belief elicitation; see, however, Offerman et al. (2009) and Hossain and Okui (2013) for alternatives.

ω_N at the truth-telling equilibrium. First, with n infinite, the proportion of a signal in the population naturally equates the probability to get that signal. Second, the agent's own signal has no impact on the asset value. The third simplification is related to the trading condition in step 4 of the Bayesian market definition. That someone else is accepting to buy or to sell at the same price is still important but the information that someone *could* make such a choice becomes trivial in an infinite group. There will always be at least one other expert receiving the same signal and one with the opposite signal. Moreover, for any nondegenerate proportion of agents endorsing each signal, there will also always be someone being offered any possible price. Hence, trade happening does not bring more information about the signal distribution than s_i does, unlike in a finite group of agents.

With the three simplifications in mind, we first address the incentive problem by the following proposition.

Proposition 1. *Let $\langle \omega_Y, \omega_N, r \rangle$ be a signal technology and n infinite. In the game induced by the Bayesian market, truth-telling is a Bayesian Nash equilibrium in which agents' betting strategies are such that:*
(i) agents whose signal is 1 buy the asset if and only if $p_i \leq \bar{\omega}_1$;
(ii) agents whose signal is 0 sell the asset if and only if $p_i \geq \bar{\omega}_0$.

Proof. The main result in Baillon (2017) is essentially unaffected by the introduction of individualized prices. To get an intuition for the result, we can inspect Equations (3) and (4). It is immediate that $\overline{\omega}_0 < \overline{\omega}_1$ since $r_0 < r_1$ and $\omega_N < \omega_Y$. Thus, signal-1 agents expect more signal-1 agents than signal-0 agents do. Consider then agent i with $s_i = 1$ and assume all other agents are telling the truth, such that the asset value v equals the true share of signal-1 agents in the population. Agent i expects v to be $\overline{\omega}_1$. For p_i less than $\overline{\omega}_1$, agent i will be willing to buy the asset. Agent i also knows that no one would buy it at a higher price (so i has no reason to pretend to be a seller) but that some agents will be willing to sell at prices between $\overline{\omega}_0$ and $\overline{\omega}_1$. For this price range, agent i foresees a profit and has the incentives to endorse

$e_i = 1$ to become a buyer. Outside this range, no trade will go through. The case $s_i = 0$ is symmetric. □

Bayesian markets avoid beauty contest effects. First, Proposition 1 shows that agents have *no* incentives to unilaterally deviate from the truth-telling strategy to report what they expect the majority endorsement will be. Second, minority agents endorsing what they expect the majority endorsement to be leads to a dominated equilibrium. Without loss of generality, imagine that $\overline{\omega}_1 < 0.5$, i.e, signal-1 agents expect to be a minority. Since $\overline{\omega}_0 < \overline{\omega}_1$, signal-0 agents expect to be an even larger majority than signal-1 agents believe. In other words, it cannot be that both groups believe they are a minority.[7] Hence, if both minority and majority agents use the strategy to endorse what they believe the majority endorsement will be, they all endorse the same state and no trade occurs (according to Step 4 of the definition of a Bayesian market). Profits are 0 for everyone, which is dominated by the strictly positive expected profits of the truth-telling equilibrium (Baillon, 2017).

The fact that agents trade an asset whose value they disagree on may raise the question why the no-trade theorem (Milgrom and Stokey, 1982) is not applicable here. The reason lies in the role of the center: For a trade to go through, it is sufficient that there exists a single agent who was willing to take the opposite bet. The center will verify this condition for each individual bettor, without providing further information about who the agent with the opposite bet is. Since $0 < \omega_N < \omega_Y < 1$, agents already know that there must be at least one disagreeing agent and thus the occurence of trade does not provide further information about the actual ω. Since trades are facilitated by the center,[8] the agents remain uncertain about the share of other agents disagreeing with them, which makes our setting different to the settings in Aumann (1976) or Milgrom and Stokey (1982) in which disagreement

[7] They may both believe to be the majority and in that case, endorsing what they expect the majority will endorse is the truth-telling equilibrium.

[8] The center will typically incur a loss from this role. The mechanism is thus not budget-balanced.

is impossible.

In the following proposition, we consider the aggregation problem and derive what conclusions the center can draw in the truth-telling equilibrium.

Proposition 2. *If n is infinite and the Bayesian market is at the truth-telling equilibrium, at least one agent has a positive payoff, all those with positive payoffs have endorsed the actual state, and all those with negative payoffs have endorsed the opposite state.*

Proof. At the truth-telling equilibrium, the settlement value v is ω_N in state N and ω_Y in state Y. And according to Proposition 1, trades only occur for prices in the range $[\overline{\omega}_0, \overline{\omega}_1]$. Hence agents' payoffs, defined in Equations (5)-(6), can be simplified as

$$\pi_1(v, p_i) = -\pi_0(v, p_i) = \begin{cases} \omega - p_i & \text{if } p_i \in [\overline{\omega}_0, \overline{\omega}_1]; \\ 0 & \text{otherwise.} \end{cases}$$

Notice that Equations (3)-(4) imply

$$0 < \omega_N < \overline{\omega}_0 < \overline{\omega}_1 < \omega_Y < 1. \tag{7}$$

In state Y, when trade occurs, signal-1 agents pay less than $\overline{\omega}_1$ and therefore less than the settlement value ω_Y. They make a profit while sellers (signal-0 agents) sell the asset at a price too low. The opposite applies in state N. Hence, the center, seeing that agents endorsing Y make a profit, can conclude that we are indeed in state Y, even though the state itself is not directly observable. Sellers making a profit indicates state N. □

At the truth-telling equilibrium and under the actual state of nature S, the average payoff for agents with the same signal s is equal to the expected payoff for agents with that signal:

$$\begin{aligned} \pi_1^Y &\equiv E_p[\pi_1(v, p) \,|\, Y] \\ &= E_p[\pi_1(\omega_Y, p)] = \int_{\overline{\omega}_0}^{\overline{\omega}_1} (\omega_Y - p) dp = \frac{1}{2}\left[(\omega_Y - \overline{\omega}_0)^2 - (\omega_Y - \overline{\omega}_1)^2\right] \\ &= -\pi_0^Y \equiv -E_p[\pi_0(v, p) \,|\, Y]; \end{aligned} \tag{8}$$

$$\begin{aligned}\pi_0^N &\equiv E_p[\pi_0(v,p)\,|\,N] \\ &= E_p[\pi_0(\omega_N,p)] = \int_{\bar{\omega}_0}^{\bar{\omega}_1}(p-\omega_N)dp = \frac{1}{2}\left[(\bar{\omega}_1-\omega_N)^2-(\bar{\omega}_0-\omega_N)^2\right] \qquad (9)\\ &= -\pi_1^N \equiv -E_p[\pi_1(v,p)\,|\,N].\end{aligned}$$

Under state Y, $\pi_1^Y > 0$ and $\pi_0^Y < 0$; and under state N, $\pi_1^N < 0$ and $\pi_0^N > 0$.

The value $\bar{\omega}_s$ is the prediction of the proportion of signal 1 in the population by agents with signal s. Hence, $\omega_S - \bar{\omega}_s$ is the prediction error of signal-s agents when S is the actual state of nature (note that this error can be positive or negative). The average payoff of signal-s agents are therefore half the difference between the squared prediction error of agents with signal $1-s$ and their own squared prediction error. Agents endorsing the actual state of nature are better able to guess the signal distribution in the population, and therefore, the opinions of others. Bayesian markets favor them and allow them to make a profit.

From observing endorsements and trades at the truth-telling equilibrium, the center can infer the whole signal technology and even beliefs. Sellers accept to sell from $\bar{\omega}_0$ onward and buyers to buy up to $\bar{\omega}_1$. Moreover, in state Y, ω gives ω_Y and ω_N is given by the formula $\omega_N = \frac{\bar{\omega}_0(\omega_Y-\bar{\omega}_1)}{\omega_Y-((1-\omega_Y)\bar{\omega}_0+\omega_Y\bar{\omega}_1)}$.[9] Equations (1) and (2) then allows us to obtain probabilistic beliefs r_1 and r_0, without having to ask how confident agents are about their endorsement.[10]

2.4. Algorithms for empirical data

Proposition 2 concerns limit behavior of perfectly rational agents. In perfect conditions, all agents endorsing the actual (opposite) state have a nonnegative (non-

[9] In state N, ω gives ω_N and ω_Y is given by the formula $\omega_Y = \frac{\bar{\omega}_0(\bar{\omega}_1-\omega_N)}{((1-\omega_N)\bar{\omega}_0+\omega_N\bar{\omega}_1)-\omega_N}$.

[10] Having r_1 and r_0 allow us to go beyond binary predictions; we can now aggregate agents' probabilistic beliefs of the actual state. For example, the simplistic way is to use the average: $\omega r_1 + (1-\omega)r_0 = \frac{(\omega\bar{\omega}_1+(1-\omega)\bar{\omega}_0)-\omega_N}{\omega_Y-\omega_N}$. We can also use more advanced methods in the probability forecast aggregation literature, e.g., various ways of extremizing the average (Ranjan and Gneiting, 2010; Satopää et al., 2014; Baron et al., 2014). Methods adapted to probabilistic forecasts have been recently proposed by Wilkening et al. (2022), Peker (2022), and Palley and Satopää (2023).

positive) payoff, and at least one agent will have a positive payoff. In practical implementation, a small group may lead to no trade. Furthermore, agents may make mistakes when endorsing a state or when deciding to trade. In the presence of noise, agents not endorsing the actual state may still make a profit. We discuss two algorithms which can be used empirically by the center to find the actual state in those non-ideal situations.

The simpler algorithm computes the payoff of each agent and compares the average payoff of the sellers to that of the buyers. The algorithm picks the side with the higher average payoff and tosses a coin if no trade occurred. We call this algorithm the *naive follow-the money algorithm* (nFTM). nFTM is able to accommodate some moderate noise in agents' behavior but does not solve the no-trade issue.

In pilot studies, we found that the presence of substantial noise hampered the accuracy of nFTM. We therefore developed a more elaborate algorithm, fitting supply and demand curves with logistic curves. For simplicity, we refer to this algorithm as the *follow-the money algorithm*, or FTM.[11] With F the logistic function, the FTM first estimates $\hat{\omega}_1$ and $\hat{\omega}_0$ (which can be interpreted as the reservation prices for an infinite group at the truth-telling equilibrium) from

$$Prob(d_i = 1 \,|\, p, e_i) = \begin{cases} F(\beta(p - \hat{\omega}_1)) & \text{if } e_i = 1 \\ F(\beta(\hat{\omega}_0 - p)) & \text{if } e_i = 0 \end{cases} \tag{10}$$

imposing $\hat{\omega}_0 \leq \hat{\omega}_1$. The logistic function has the following properties that make it suitable for our purpose: for prices that are lower for the buyers or higher for the sellers than their respective reservation prices, the probability of taking the bet is higher than 0.5 and increasing with the distance to reservation price; for prices which equal the reservation prices, there is a 0.5 chance of taking the bet; for prices too high for the buyers or too low for the sellers, the probability of taking the bet is lower than 0.5 and decreasing with the distance to reservation price. Parameter

[11] In our preregistration, we refer to nFTM and FTM as FTM-A and FTM-L, for "average" and "logistic", instead.

β captures the level of noise/imprecision (sensitivity towards the distance between the price and reservation price) and is assumed to be the same for sellers and buyers (for parsimony). FTM then computes the expected payoffs for buyers and sellers for an infinite group using Equations (8) or (9), substituting $\bar{\omega}_1$ and $\bar{\omega}_0$ with estimated reservation prices $\hat{\omega}_1$ and $\hat{\omega}_0$, and ω with the proportion of endorsements 1, and picks the side with a positive expected payoff.

3. Experimental design

3.1. Stimuli

We conducted an experiment with abstract tasks (urns and balls) ensuring that the theoretical assumptions were satisfied. We considered groups of $n = 200$ agents. In each task, the participants of the experiment were presented with two urns, as depicted in Figure 1. Urns Left and Right represent the two states of nature, N and Y respectively. Participants were told that one of the two urns was selected randomly ($r = 0.5$) and that each of the 200 participants of a group would get one ball from that urn. Denoting a yellow ball $s_i = 1$ and a blue ball $s_i = 0$, Urn Left would give $\omega_N = 0.10$ and Urn Right $\omega_Y = 0.40$ in this particular example. Urn Right always contains more yellow balls than Urn Left. Thus Urn Right is state of nature Y and Urn Left is state of nature N.

There were 30 tasks with ω_N ranging from 0.05 to 0.75 and ω_Y from 0.25 to 0.95, spanning the unit interval in a systematic way. In twelve tasks, both urns had a minority of yellow balls, i.e., $\omega_N < \omega_Y < 0.5$. Another set of twelve tasks mirrored them such that $\omega_Y > \omega_N > 0.5$, and in six tasks the majority would always guess the correct state of nature ($\omega_Y > 0.5 > \omega_N$). Table 3 in Online Appendix A lists all the task parameters. The number of yellow balls differs across states of nature by a minimum of 40 and a maximum of 60. Larger differences would mean that the signal technology discriminates very well between state of nature and the majority (as well

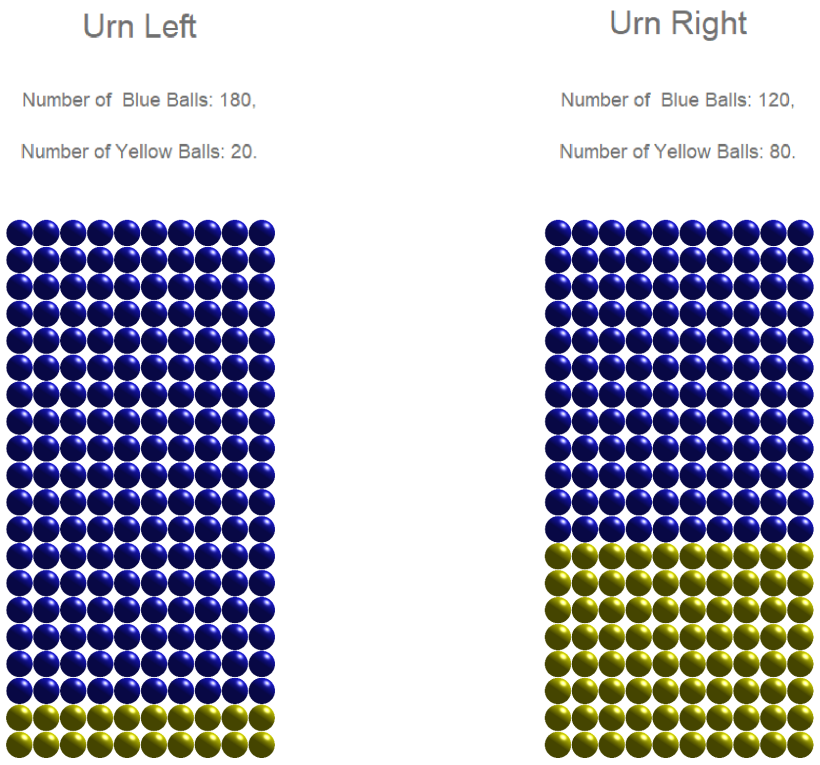


Figure 1: Experimental task setting (an example of Task 6)

as FTM) would be right most of the time. By contrast, smaller differences would imply very narrow trading intervals and it could be that none of the 200 participants of a group gets a price in that range.

In each task, the participants were presented with the urns (as in Figure 1) and asked to press a button to draw their ball. Once the color of their ball was revealed, they were asked to guess which urn the ball comes from (i.e. to endorse a state). The next question differed between two experimental treatments, FTM and SPA.

In the FTM treatment, we implemented the betting mechanism of the Bayesian markets. In Figure 2 for instance, participants were asked whether they were willing to bet that the number of participants guessing Urn Right (i.e. endorsing Y) was at least 130, i.e., whether they were willing to pay $p = 0.65$ for v, which is equal to the sample proportion ω. For the sake of symmetry, participants guessing Urn Left were asked whether they would bet that the number of participants guessing

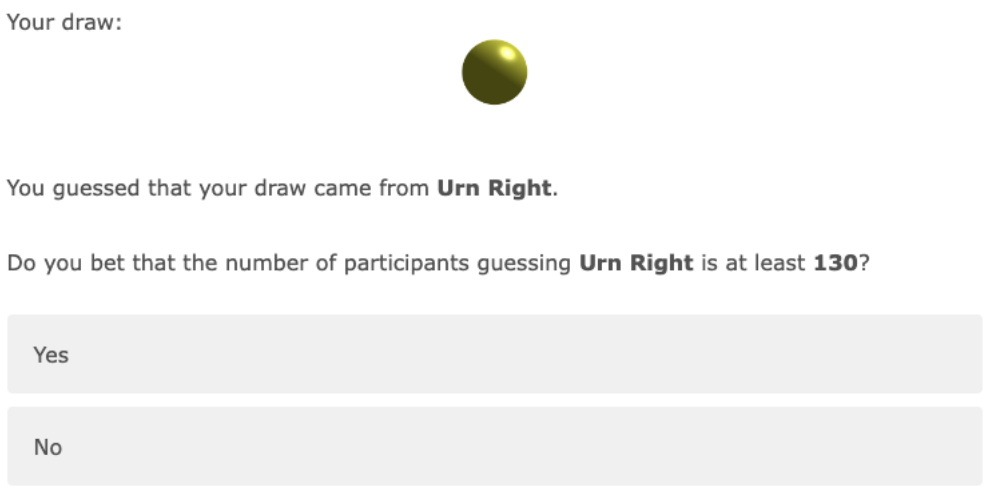


Figure 2: Screenshot of the FTM treatment

Urn Left would be at least 70, i.e. whether they were willing to accept $p = 0.65$ for v. Payment was explained in a training preceding the experiment. The participants were told that the number (130 in this example) was random and that their payment would be the actual number of Urn Right guesses minus that number if someone took the opposite bet (betting that at least 70 or more participants would guess Urn Left). It would be 0 otherwise.

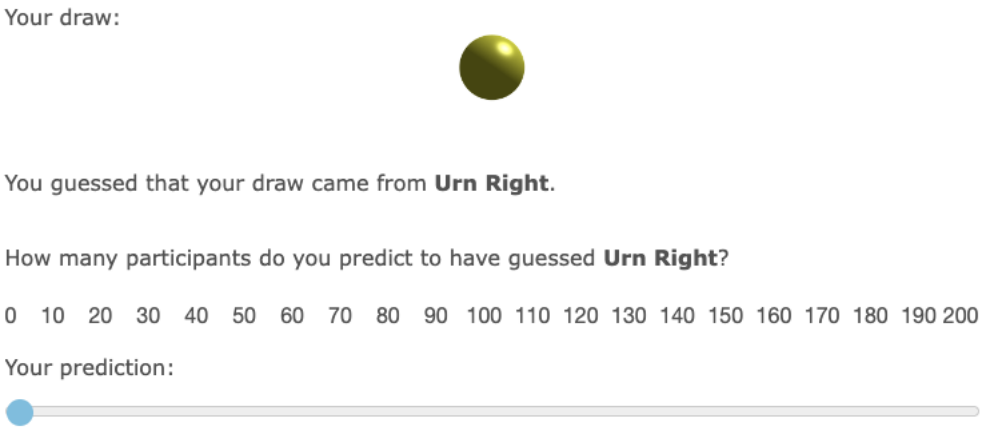


Figure 3: Screenshot of SPA

In the SPA treatment, we followed the approach of Prelec et al. (2017) and asked participants to predict the number of people who guessed the same urn as they did

(Figure 3). Prelec et al.'s (2017) algorithm first computes the average prediction across all participants and then selects the state of nature that was endorsed more often than predicted. Predictions were incentivized using the quadratic scoring rule. Participants received $400-\frac{x^2}{100}$, with x the difference between their prediction and the actual number of guesses. There were no incentives for making correct endorsements.

In none of the treatments was payment directly based on the task parameters. Even though we, the experimenters, knew them, we aimed to mimic situations in which no one knows the actual state of nature and in which the center (paying the agents) does not even know the signal technology.

3.2. Deviations from Section 2

The implementation of the Bayesian market in our experiment differs from the Bayesian market mechanism proposed in section 2 in two ways. First, n is, trivially, not infinite. Hence, there is no guarantee that the incentive properties established in Proposition 1 are preserved. We performed extra checks in Online Appendix A to make sure that truth-telling equilibrium exists for parameter values used in our experiment and for $n = 200$. Online Appendix A shows that reservation prices may differ from $\overline{\omega}_0$ and $\overline{\omega}_1$ (the reservation prices for an infinite sample). Table 3 provides the reservation prices for $n = 200$, denoted p_0^* and p_1^* for sellers and buyers respectively. Second, the draws from the urn (i.e., the signals) were made without replacement. Beyond simplifying calculations for respondents, it also implies that the settlement value could only be ω_Y or ω_N if participants endorse the state corresponding to their signal. Consequently, if they do follow the strategy of Proposition 1, the aggregation properties established in Propositions 2 are preserved. Hence, we make the following two observations:

Observation 1: Given the setup of the experiment, the incentive properties established in Proposition 1 are preserved with reservation prices p_0^* and p_1^* instead of $\overline{\omega}_0$ and $\overline{\omega}_1$.

Observation 2: Given the setup of the experiment, the aggregation properties

established in Propositions 2 are preserved.

3.3. Implementation

The experiment was preregistered (https://osf.io/cf8bk/) and conducted on Prolific between July 24 and August 9, 2022, with 828 participants in the FTM treatment and 822 in the SPA treatment. They were all US students. We restricted participation to students for their probable familiarity with abstract tasks as those used in our experiment. Participants watched a short video explaining the experimental tasks and then went through five training rounds where they received feedback about their payments and how these payments were calculated (see Online Appendix C for details). We split the 30 tasks into two sets of 15. After the training, each participant completed one of the two sets, with the task order being randomized within that set at the participant level. There was no feedback after the tasks. Payment, described in the next paragraph, occurred once all participants had completed the experiment.

Participants received a fixed reward of £1.5 and a bonus of up to £3.[12] All amounts (prices, bets, scores) were presented in tokens. The bonus in pounds was the number of tokens divided by 2,000. In the FTM treatment, participants could (in theory) win or lose up to 200 tokens in each task. Hence, they were endowed with 200 tokens for each task to avoid net losses at the end of the experiment. In the SPA treatment, the quadratic score was also expressed in tokens. It was equivalent to endowing them with 400 tokens and imposing a quadratic loss ranging from 0 to 400. In both treatments, the final number of tokens was naturally bounded by 0 and 6,000. This allowed us to recruit participants with the same information about bonus ranges. However, the average bonus was likely to be lower for the FTM treatment than for the SPA treatment ex ante and, in fact, it was ex post (SPA £2.86, FTM £1.64).

To compute the bonus of a participant in a given task after the end of the

[12]Prolific required payments in pounds.

experiment, we randomly selected a state of nature[13] and 200 participants such that the group (including this particular participant) had the exact combination of signals shown in the task. In other words, participants were not assigned to a given group ex ante. Instead, we constructed (random) groups matching the information provided to the participants.

4. Results

To be consistent, we report data and results in terms of our theoretical setting. In particular, a yellow ball is signal 1 ($s_i = 1$) and a blue ball is signal 0 ($s_i = 0$). A participant guessing Urn Right is endorsing state of nature Y ($e_i = 1$) and guessing Urn Left is endorsing N ($e_i = 0$).[14] We also define truth-telling as reporting $e_i = s_i$. We cannot distinguish those who did not tell the truth from those who did not update their belief correctly. Hence, we also refer to $e_i = s_i$ as reporting a Bayesian guess. The analysis was preregistered, with the exception of the exploratory subsection 4.4.

4.1. Raw data - Endorsements

According to the model, truth-telling would be a Bayesian Nash equilibrium in the FTM treatment. The empirical truth-telling rate was 90.1%.[15] About 59% of the

[13]This random selection of a state of nature resulted in 50.1% of state Y selected for bonus calculations of participants in the SPA treatment, and 49.6% for the FTM treatment. Both proportions are not significantly different from 0.5 (proportion tests: for SPA, Z-statistic= 0.216, $p = 0.829$; and for FTM, Z-statistic= -0.969, $p = 0.333$).

[14]Predictions elicited in the SPA treatment were about the number of participants guessing the same urn, but we deduct the predictions of participants endorsing N from 200 to be the predictions of number of participants endorsing Y. Bets in the FTM treatment were also expressed in terms of the number of participants guessing the same urn, but we deduct the prices in the bets for participants endorsing N from 200 to be the the prices to sell the asset whose settlement value is the number of participant endorsing Y.

[15]The empirical truth-telling rates were not significantly different for easier questions with $\omega_Y > 0.5 > \omega_N$ and for other questions with $\omega_Y > \omega_N > 0.5$ or $0.5 > \omega_Y > \omega_N$ (89.7% and 90.2%

participants told the truth in all 15 tasks they faced. About 23% guessed the opposite urn (or lied about their guess) 1 to 3 times out of 15. Less than 4% had a majority of lies / wrong guesses (Table 4 in Online Appendix B). The incentives provided in the SPA treatment did not make truth-telling a Bayesian Nash equilibrium, but we observed a very similar truth-telling rate (SPA: 90.0% of the cases, not significantly different from FTM, with proportion test Z-statistic$= -0.254$ and $p = 0.800$).

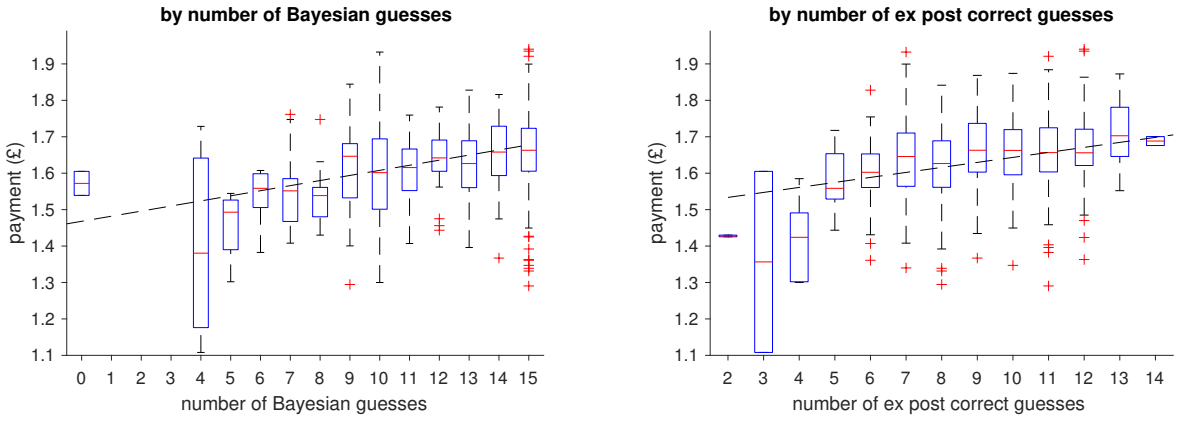


Figure 4: Payments in FTM.

The truth-telling rate of the FTM treatment was sufficiently high to reward those who correctly reported their signal and to penalize those who lied or misreported. The left panel of Figure 4 displays earnings as a function of the number of times participants told the truth. It shows a positive correlation, with a fitted line slope of 0.014 ($p < 0.001$). Participants did not get feedback during the experiment (only in the five training rounds). The figure illustrates that feedback about payment could have improved truth-telling rate by allowing participants to learn that correctly reporting their signal is rewarded. It further shows that in future experiments one can announce in the instructions that a previous study showed that participants who tell the truth more often can earn more in such a setting.

So far, we studied what the raw data told us about participants' strategic behavior, illustrating the incentive properties of Bayesian markets (Proposition 1).

respectively; proportion test Z-statistic$= 0.828$ and $p = 0.408$).

To illustrate the aggregation properties (Proposition 2), we can check whether accurately guessing the selected urn led to higher earnings in our experiment. The prediction is supported by the right panel of Figure 4, which is a box plot of earnings as a function of the number of times participants guessed the actual state. The fitted line slope is 0.014 ($p < 0.001$). Thus, Bayesian markets reward expertise. While in our experiment, this expertise is artificially created,[16] in many applications one may expect that the number of times someone guesses the actual state of the world to be influenced by a more natural notion of expertise, i.e. domain knowledge.

4.2. Raw data - Predictions and trades

If participants are Bayesian, they should report the posteriors $\overline{\omega}_0$ and $\overline{\omega}_1$ in the SPA treatment, at least if they expect everyone else to tell the truth. Figure 5 displays the average predictions as a function of theoretical posteriors for both type of guess. Predictions are very close to Bayesianism for $\overline{\omega}_0 < 0.5$ and $\overline{\omega}_1 > 0.5$. Interestingly, participants seemed to have much more difficulty to predict that a majority of people would guess Y when they themselves guess N or that only a minority would guess Y when they themselves guess Y. The SPA uses the average prediction across both guesses, which mitigates this issue.

We do not have participants' predictions in the FTM treatment but we can compare the participants' decisions d_i to the theoretical predictions. Table 1 compares the theoretical and empirical proportions of $d_i = 1$ (the willingness to buy / to sell) for five price intervals, defined by ω_N, p_0^*, p_1^*, and ω_Y. Buyers should be willing to pay at most p_1^* and sellers willing to accept not less than p_0^*. If participants do not compute the Bayesian posterior but use ω_Y and ω_N instead, i.e. the distribution of balls of the urn they guessed, buyers would be willing to pay at most ω_Y and sellers willing to accept not less than ω_N. If they were extremely risk averse, buyers would be willing to pay at most ω_N and sellers willing to accept not less than ω_Y.

The empirical willingness to sell was increasing with price and the empirical

[16] It consists of receiving informative signals, in combination with a truth-telling strategy.

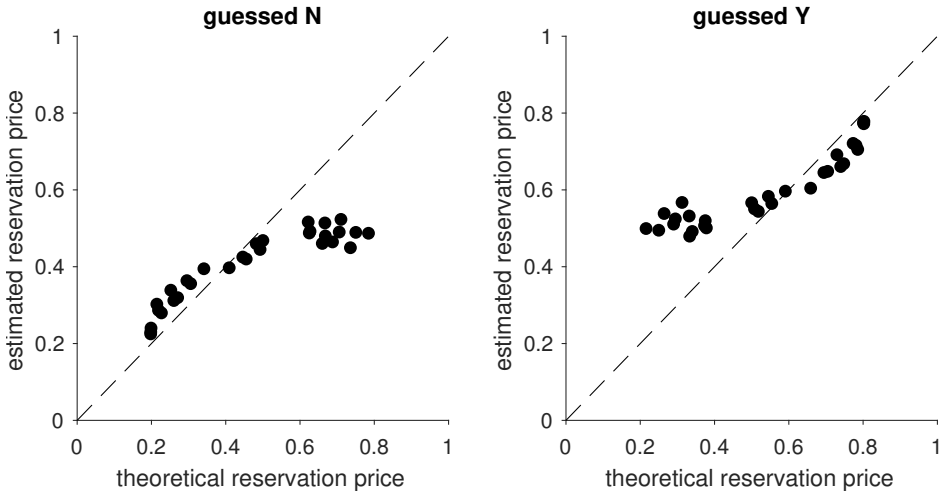


Figure 5: Theoretical $\overline{\omega}_0$ and $\overline{\omega}_1$ vs. average predictions in SPA.

Table 1: Theoretical and empirical bet acceptance (in %) and average payoffs (in tokens) in the FTM treatment by price interval

	$p \in$	$[0, \omega_N)$	$[\omega_N, p_0^*)$	$[p_0^*, p_1^*]$	$(p_1^*, \omega_Y]$	$(\omega_Y, 1]$
guessed N (seller)	theo. acceptance	0%	0%	100%	100%	100%
	emp. acceptance	29.4%	50.3%	70.1%	77.7%	92.3%
	average payoff	-53.0	-12.6	5.3	21.7	53.6
guessed Y (buyer)	theo. acceptance	100%	100%	100%	0%	0%
	emp. acceptance	91.3%	80.3%	67.9%	48.1%	28.6%
	average payoff	53.1	21.4	3.9	-12.0	-51.9

willingness to buy was decreasing, as predicted in the truth-telling equilibrium. However, for several participants the acceptance and rejection ranges of bets were not consistent with the equilibrium prediction. About 30% of bets that are losing for sure under truth-telling were accepted and about 30% of bets that are winning for sure under truth-telling were rejected (see leftmost column of the seller row and rightmost column of the buyer row). In total, there was a clear tendency to bet much more than predicted by equilibrium play.

Table 1 also reports the average payoffs of the participants for each price inter-

val. The results confirm that participants who accepted bets that would have been losing for sure if everyone else had told the truth, still bore a loss on average in our experiment. Overall, trading decisions were noisy and substantially deviated from the theoretical predictions. This underlines that the performance of the FTM algorithms will depend on their ability to recover aggregate reservation prices from the noisy trades.

4.3. Accuracy comparison

The final part of the analysis aims to compare accuracy of the various methods. We want to assess the ability of the majority rule, SPA, and FTM algorithms to identify the actual state of nature using the participants' answers.

To make full use of the answers of all respondents who provided answers to a task, we ran 1,000 simulations for each task, state of nature, and treatment, randomly making groups of 200 participants. For instance, consider one of the simulations for the task described in Figure 1 with $\omega_Y = 0.40$ and $\omega_N = 0.10$ (Task 6), state of nature Y, and the FTM treatment. We randomly composed a group of 200 FTM participants, such that exactly 80 of them had gotten $s_i = 1$. We then use the answers from the 200 participants to determine the state using majority rule, nFTM, and FTM. Similarly, we randomly composed 1,000 groups of 200 SPA participants in the same way to determine the state using majority rule and SPA. Repeating the same procedures for each of the 30 tasks and two possible states of nature, we obtained 60 accuracy rates for each method. Table 2 summarizes the average accuracy rates for each algorithm and for the majority rule. We conducted Wilcoxon tests to test for differences.

Table 2 distinguishes two cases. If $\omega_N < 0.5 < \omega_Y$ (top row), then the majority rule should determine the actual state all the time. In the other cases (bottom row), the majority rule finds the actual state 50% of the time, by pure chance. Our results are consistent with these predictions (see columns 'majority rule'), both for the data from the SPA treatment and for the data of the FTM treatment. The SPA, our

Table 2: Average accuracy rates from simulations

cluster of questions	majority rule		SPA	FTM	
	SPA data	FTM data		*nFTM*	*FTM*
$\omega_Y > 0.5 > \omega_N$	100.0%	100.0%	100.0%	98.4%	99.7%
$\omega_Y > \omega_N > 0.5$ or $0.5 > \omega_Y > \omega_N$	50.0%	50.2%	77.6%	55.1%	73.8%

benchmark, should always identify the actual state if participants were Bayesian and reporting truthfully all the time. In spite of non-Bayesian answers and noise, SPA performed as well as majority when $\omega_N < 0.5 < \omega_Y$, and substantially improved upon majority when following the majority is equivalent to tossing a coin (Wilcoxon signed rank test $p < 0.001$). In that case, the average accuracy increased by 27.6 percentage points (pp).

Computing average payoffs on Bayesian markets, as our nFTM algorithm does, led to worse results than SPA, whether $\omega_N < 0.5 < \omega_Y$ (Wilcoxon signed rank test $p = 0.016$) or not (Wilcoxon signed rank test $p < 0.001$). nFTM is highly sensitive to noise and we noticed earlier that our data were clearly noisy. To account for noise, FTM fits logistic supply and demand curves on the buy and sell decisions and only then computes expected payoffs. FTM substantially improved upon nFTM (Wilcoxon signed rank test $p < 0.001$), especially when $\omega_Y > \omega_N > 0.5$ or $0.5 > \omega_Y > \omega_N$ (Wilcoxon signed rank test $p < 0.001$), with an increase of 18.7pp. It yielded results that were not significantly different from SPA (Wilcoxon signed rank tests $p = 0.272$), especially when $\omega_Y > \omega_N > 0.5$ or $0.5 > \omega_Y > \omega_N$ (Wilcoxon signed rank tests $p = 0.395$). Interestingly, it gave results comparable to SPA with less information. SPA uses, as input, an endorsement and a prediction (number between 0 and 1), directly asking participants for $\overline{\omega}_0$ and $\overline{\omega}_1$. FTM uses an endorsement and a trade decision, which is binary. FTM compensates the information loss by using (simple) econometric techniques to recover reservation prices. FTM significantly improved upon majority in general (Wilcoxon signed rank tests $p < 0.001$), especially when majority rule assumption does not hold (Wilcoxon signed rank tests $p < 0.001$).

4.4. Exploratory analysis

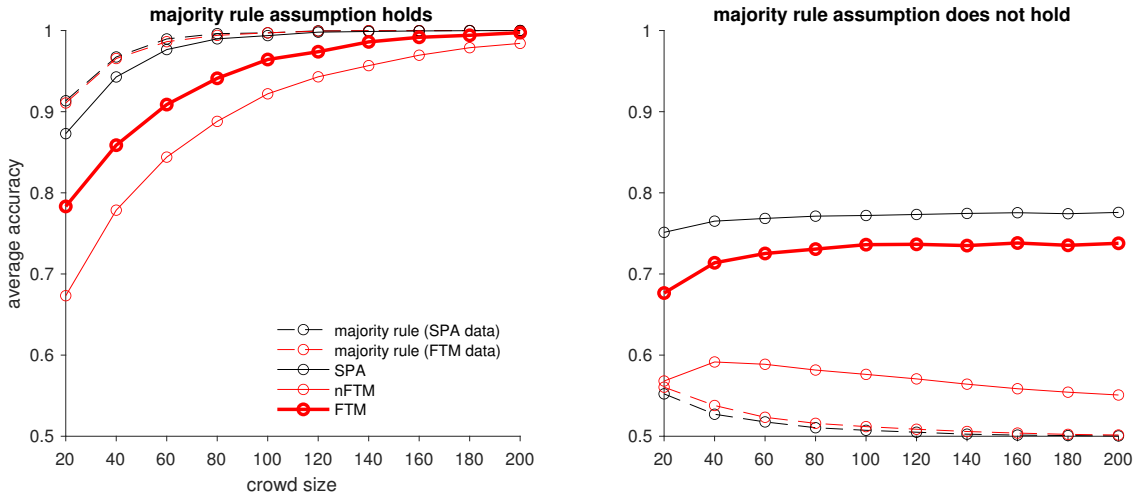


Figure 6: Accuracy comparisons for different group sizes.

The accuracy analysis so far was based on groups of 200 participants. We can also study how sensitive results are to group sizes. We replicated the analysis (with 1,000 simulations for each combination of method, task, and state of nature) for various group sizes ranging from 20 to 180. This part of the analysis was not preregistered, which is why we call it exploratory. Figure 6 depicts the accuracy rates as a function of group size. FTM is more sensitive to group size than SPA. This is because the FTM profits from additional observations in order to better estimate supply and demand curves. In the left panel, when $\omega_N < 0.5 < \omega_Y$, accuracy of FTM increases from around 78% for groups of 20 to almost 100% for groups of 200. SPA performs better for small groups. This is in line with the general tendency of prediction markets to have accuracy increasing with market thickness. In the right panel ($\omega_Y > \omega_N > 0.5$ or $0.5 > \omega_Y > \omega_N$), SPA is very stable, with accuracy rates between 75% and 78%, while the accuracy of FTM increases from 68% for groups of 20 to 74% for groups of 200. These results are not surprising, knowing that, for groups of 20, FTM has to determine reservations prices from very few binary decisions (buying or selling). There may be as little as one buyer or one

seller in some groups even if everybody reports truthfully.

5. Discussion

In our experiment, each agent received an endowment to avoid losses. Even without providing an endowment, agents can expect a strictly positive payoff (Baillon, 2017), which can motivate them to participate. The center, who plays the role of the market maker, subsidizes the market and acts as an intermediary between the agents, who do not trade with each other. Absent this intermediary role of the center, agents would infer others' signals from their willingness to buy or sell. Similar to the classical reasoning in Aumann (1976) and Milgrom and Stokey (1982), they would then agree on the state, leaving no room for trade based on disagreement.

To communicate the same information to all potential participants, we fixed the bonus range from £0 to £3. The SPA was more expensive (SPA £2.86, FTM £1.64). If anything, the SPA participants, with an endowment of 400 tokens and a quadratic loss should have been more motivated than the FTM participants. The SPA treatment only incentivized predictions, not truthful endorsement. The latter could have been done using the Bayesian truth serum of Prelec (2004) but the payoff rule is difficult to explain to participants. Experiments that have been using this truth serum did not explain the payoff function in detail, but rather used an "intimidation method", telling participants it is in their interest to tell the truth. We refrained from such an approach, and instead included instructions and training to explain our payoff rules. An alternative for future research is to incentivize the SPA using choice-matching (Cvitanić et al., 2019), which elicits predictions and endorsements with a simpler payment formula.

Regarding our theoretical model, four restrictive assumptions warrant some further discussion. First, recall that that we treated strategies as maps from signals to endorsements, such that agents could not make their endorsements depend on any other event or randomization device, and did not allow asymmetric strategies.

Second, we allowed no communication between agents. Third, we only considered a binary underlying state space and fourth, our market setting is a one-period, static setting. We discuss each of those in turn.

It is important that agents cannot coordinate on events other than type realizations. Among the remaining symmetric equilibria, the truth-telling equilibrium is (ex-ante) Pareto optimal.[17] It is behaviorally plausible, as conjectured by Baillon (2017), that truth-telling is focal and, in our experiment, there was indeed little evidence of agents trying to find a reverse strategy. More than half of the participants consistently told the truth and a negligible share of participants chose to systematically misstate their type (see Table 4 in Online Appendix B). Without the aforementioned restriction, agents could try to coordinate on some other signal, in which the probability of receiving a 0-signal in state Y and a 1-signal in state N is very low. Then, a small number of agents will make a loss of (almost) 1, and a large share of agents will make a profit of (almost) 1. In expectation, all agents thus have a high expected payoff. Note that this coordination does not only require mere communication among respondents but also some credible randomization device. To avoid such coordinated attacks, the center should make it an active feature of design that market participants are (at least partially) anonymous, as is the case in our experiment.

As suggested by the previous paragraph, our approach cannot be used if there is public discussion of private signals or if agents can form coalitions. If it is possible to bring all experts together, other approaches to the aggregation problem have been proposed in the literature, such as the Delphi method developed in the 1950s at the Rand Corporation (Okoli and Pawlowski, 2004). These approaches do not solve the incentive problem though. In our setting, experts do not have other incentives than those we provide to hide or manipulate their private information. The literature on

[17] To see this, note first that it is pay-off equivalent to the "reverse" equilibrium in which everyone endorses the state that they believe to be less likely to be the actual one. The two other equilibria have universal endorsement of either state Y or of state N and thus obviously lead to a universal payoff of zero since there is never any trade.

committee decisions studies how agents may agree to share their private signals with each other in order to look united if their reputation is at stake (Visser and Swank, 2007; Swank et al., 2008).

We considered a binary state space. If the state space is non-binary, one may organize several Bayesian markets, with different agents. Consider three states A, B, and C, and assume agents can choose which state they would like a signal about (e.g., an agent can design an experiment testing whether we are in state A or not-A). The center can assign agents to markets, inform them about which state their market will be about, let them run their experiment for that state, and then organize the Bayesian markets.

Bayesian markets and their aggregation properties can further be translated to a setting in which a market is run continuously. Suppose that there are T periods and that for each $t = 1, ..., T$, a Bayesian market is set up to trade on an asset v_t that represents the share of buyers in the Bayesian market at time t. All of these markets are only settled at the final period T, so that in particular agents do not learn the value of the assets. At each t, the incentive and aggregation properties of Bayesian markets are not affected by the markets in other periods. A continuous market can sometimes be advantageous for the center: Suppose for instance that the signal technology is constant across all periods, but that the actual state S (and therefore ω_S) may vary with t. Once the center has found a market-clearing price p^* (i.e. a price at which each agent is willing to either buy or sell the asset), this price can be chosen for any subsequent period. Since the signal technology is the same, this price will now lead to trade in each period, thereby reducing the payoff-uncertainty faced by the agents. Then, the center can make inferences about the change of the state over time by computing which side would make a profit if the market was settled. Furthermore, if the signal technology is not fixed, this will be reflected in the buying and selling decisions of the agents, and henceforth the center can detect such changes.

Noise in our results show that there is still room for simplifying belief elicitation,

for instance using frequency formulations. Such an approach, when states of nature are observable, has been proposed by Schlag and Tremewan (2021). They show how this approach dominates the method proposed by Karni (2009), which has a random component from which Step 2 of Bayesian markets is inspired.

The literature on the wisdom of crowds started with the intuition that asking many people may be better than relying on a few experts. Some have raised doubts on the mere possibility to "chase the experts" within a group (Larrick and Soll, 2006). However, there is still value to ask large groups of experts. DellaVigna and Pope (2017) found that the aggregated opinion of academic experts is closer to experimental results than estimates based on a meta-analysis of previous empirical findings. In a follow-up study, DellaVigna and Pope (2018) also showed that academic experts better predict than non-experts, even though degrees of expertise (among experts) such as academic rank or citations do not correlate with performance. Aggregating the opinions of very large group of experts becomes more and more common, for instance the International Panel on Climate Change or surveys of economists and financial specialists about future economic indicators.

6. Conclusion

Mechanisms such as prediction markets and scoring rules can incentivize truthful signal reporting and aggregate the reports. They are not applicable though if the state of the world is not objectively observable. In such a case, payoffs cannot be state-contingent, creating an incentive problem. Furthermore, in many plausible situations, one may prefer not to rely on the majority opinion, at least if experts themselves, aware of the signal structure, would not. We demonstrated theoretically and empirically how to solve both the incentive and the aggregation problem at once. Agents bet on others' endorsement and their payoffs reveal the state of nature. When implemented in a large online experiment, our follow-the-money approach performed as well as a recent alternative, the surprisingly popular algorithm, with

less information from participants.

Conflict of interest statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Andersen, S., Fountain, J., Harrison, G. W., and Rutström, E. E. (2014). Estimating subjective probabilities. *Journal of Risk and Uncertainty*, 48:207–229.

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics*, pages 1236–1239.

Baillon, A. (2017). Bayesian markets to elicit private information. *Proceedings of the National Academy of Sciences*, 114(30):7958–7962.

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., and Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 11(2):133–145.

Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232.

Cvitanić, J., Prelec, D., Riley, B., and Tereick, B. (2019). Honesty via choice-matching. *American Economic Review: Insights*, 1(2):179–92.

DellaVigna, S. and Pope, D. (2017). What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069.

DellaVigna, S. and Pope, D. (2018). Predicting experimental results: Who knows what? *Journal of Political Economy*, 126(6):2410–2456.

Enke, B. and Zimmermann, F. (2019). Correlation neglect in belief formation. *The Review of Economic Studies*, 86(1):313–332.

Hossain, T. and Okui, R. (2013). The binarized scoring rule. *Review of Economic Studies*, 80(3):984–1001.

Jaffray, J.-Y. and Karni, E. (1999). Elicitation of subjective probabilities when the initial endowment is unobservable. *Journal of Risk and Uncertainty*, 18:5–20.

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica*, 77(2):603–606.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., tpn Bahnk, Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., Rdei, A. C., Cai, H., Cambier, F., Cantarero, K., Carmichael, C. L., Ceric, F., Chandler, J., Chang, J.-H., Chatard, A., Chen, E. E., Cheong, W., Cicero, D. C., Coen, S., Coleman, J. A., Collisson, B., Conway, M. A., Corker, K. S., Curran, P. G., Cushman, F., Dagona, Z. K., Dalgar, I., Rosa, A. D., Davis, W. E., de Bruijn, M., Schutter, L. D., Devos, T., de Vries, M., Doulu, C., Dozo, N., Dukes, K. N., Dunham, Y., Durrheim, K., Ebersole, C. R., Edlund, J. E., Eller, A., English, A. S., Finck, C., Frankowska, N., ngel Freyre, M., Friedman, M., Galliani, E. M., Gandi, J. C., Ghoshal, T., Giessner, S. R., Gill, T., Gnambs, T., ngel Gmez, Gonzlez, R., Graham, J., Grahe, J. E., Grahek, I., Green, E. G. T., Hai, K., Haigh, M., Haines, E. L., Hall, M. P., Heffernan, M. E., Hicks, J. A., Houdek, P., Huntsinger, J. R., Huynh, H. P., IJzerman, H., Inbar, Y., se H. Innes-Ker, Jimnez-Leal, W., John, M.-S., Joy-Gaba, J. A., Kamilolu, R. G., Kappes, H. B., Karabati, S., Karick, H., Keller, V. N., Kende, A., Kervyn, N., Kneevi, G., Kovacs, C., Krueger, L. E., Kurapov, G., Kurtz, J., Lakens, D., Lazarevi, L. B., Levitan, C. A., Neil A. Lewis, J., Lins, S., Lipsey, N. P., Losee, J. E., Maassen, E., Maitner, A. T., Malingumu, W., Mallett, R. K., Marotta, S. A., Meedovi, J., Mena-Pacheco, F., Milfont, T. L., Morris, W. L., Murphy, S. C., Myachykov, A., Neave, N., Neijenhuijs, K., Nelson, A. J., Neto, F., Nichols, A. L., Ocampo, A., ODonnell, S. L., Oikawa, H., Oikawa, M., Ong, E., Orosz, G., Osowiecka, M., Packard, G., Prez-Snchez, R., Petrovi, B., Pilati, R., Pinter,

B., Podesta, L., Pogge, G., Pollmann, M. M. H., Rutchick, A. M., Saavedra, P., Saeri, A. K., Salomon, E., Schmidt, K., Schnbrodt, F. D., Sekerdej, M. B., Sirlop, D., Skorinko, J. L. M., Smith, M. A., Smith-Castro, V., Smolders, K. C. H. J., Sobkow, A., Sowden, W., Spachtholz, P., Srivastava, M., Steiner, T. G., Stouten, J., Street, C. N. H., Sundfelt, O. K., Szeto, S., Szumowska, E., Tang, A. C. W., Tanzer, N., Tear, M. J., Theriault, J., Thomae, M., Torres, D., Traczyk, J., Tybur, J. M., Ujhelyi, A., van Aert, R. C. M., van Assen, M. A. L. M., van der Hulst, M., van Lange, P. A. M., van t Veer, A. E., Vsquez-Echeverra, A., Vaughn, L. A., Vzquez, A., Vega, L. D., Verniers, C., Verschoor, M., Voermans, I. P. J., Vranka, M. A., Welch, C., Wichman, A. L., Williams, L. A., Wood, M., Woodzicka, J. A., Wronska, M. K., Young, L., Zelenski, J. M., Zhijia, Z., and Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490.

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

Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johanneson, M., Kirchler, M., Razen, M., Weitzel, U., Abad, D., Abudy, M. M., Adrian, T., Ait-Sahalia, Y., Akmansoy, O., Alcock, J., Alexeev, V., Aloosh, A., Amato, L., Amaya, D., Angel, J. J., Bach, A., Baidoo, E., Bakalli, G., Barbon, A., Bashchenko, O., Bindra, P. C., Bjonnes, G. H., Black, J. R., Black, B. S., Bohorquez, S., Bondarenko, O., Bos, C. S., Bosch-Rosa, C., Bouri, E., Brownlees, C. T., Calamia, A., Cao, V. N., Capelle-Blancard, G., Capera, L., Caporin, M., Carrion, A., Caskurlu, T., Chakrabarty, B., Chernov, M., Cheung, W. M. Y., Chincarini, L. B., Chordia, T., Chow, S. C., Clapham, B., Colliard, J.-E., Comerton-Forde, C., Curran, E., Dao, T., Dare, W., Davies, R. J., De Blasis, R., De Nard, G., Declerck, F., Deev, O., Degryse, H., Deku, S., Desagre, C., van Dijk, M. A., Dim, C., Dimpfl, T., Dong, Y. J., Drummond, P., Dudda, T. L., Dumitrescu, A., Dyakov, T., Dyhrberg, A. H., Dzieliski, M., Eksi, A., El Kalak, I., ter Ellen, S., Eugster, N., Evans, M. D., Farrell, M., Flez-Vias, E., Ferrara, G., Ferrouhi, E. M., Flori, A., Fluharty-Jaidee,

J., Foley, S., Fong, K. Y. L., Foucault, T., Franus, T., Franzoni, F. A., Frijns, B., Frmmel, M., Fu, S. M., Fllbrunn, S., Gan, B., Gehrig, T., Gerritsen, D., Gil-Bazo, J., Glosten, L. R., Gomez, T., Gorbenko, A., Gbilmez, U., Grammig, J., Gregoire, V., Hagstrmer, B., Hambuckers, J., Hapnes, E., Harris, J. H., Harris, L., Hartmann, S., Hasse, J.-B., Hautsch, N., He, X.-Z. T., Heath, D., Hediger, S., Hendershott, T. J., Hibbert, A. M., Hjalmarsson, E., Hoelscher, S., Hoffmann, P., Holden, C. W., Horenstein, A. R., Huang, W., Huang, D., Hurlin, C., Ivashchenko, A., Iyer, S. R., Jahanshahloo, H., Jalkh, N., Jones, C. M., Jurkatis, S., Jylha, P., Kaeck, A., Kaiser, G., Karam, A., Karmaziene, E., Kassner, B., Kaustia, M., Kazak, E., Kearney, F., van Kervel, V., Khan, S., Khomyn, M., Klein, T., Klein, O., Klos, A., Koetter, M., Krahnen, J. P., Kolokolov, A., Korajczyk, R. A., Kozhan, R., Kwan, A., Lajaunie, Q., Lam, F. Y. E. C., Lambert, M., Langlois, H., Lausen, J., Lauter, T., Leippold, M., Levin, V., Li, Y., Li, M. H., Liew, C. Y., Lindner, T., Linton, O. B., Liu, J., Liu, A., Llorente-Alvarez, J.-G., Lof, M., Lohr, A., Longstaff, F. A., Lopez-Lira, A., Mankad, S., Mano, N., Marchal, A., Martineau, C., Mazzola, F., Meloso, D. C., Mihet, R., Mohan, V., Moinas, S., Moore, D., Mu, L., Muravyev, D., Murphy, D., Neszveda, G., Neumeier, C., Nielsson, U., Nimalendran, M., Nolte, S., Nordn, L. L., O'Neill, P., Obaid, K., degaard, B. A., stberg, P., Painter, M., Palan, S., Palit, I., Park, A., Pascual Gasc, R., Pasquariello, P., Pastor, L., Patel, V., Patton, A. J., Pearson, N. D., Pelizzon, L., Pelster, M., Prignon, C., Pfiffer, C., Philip, R., Plhal, T., Prakash, P., Press, O.-A., Prodromou, T., Putnins, T. J., Raizada, G., Rakowski, D. A., Ranaldo, A., Regis, L., Reitz, S., Renault, T., Wang, R., Ren, R., Riddiough, S., Rinne, K., Rintamki, P., Riordan, R., Rittmannsberger, T., Rodrguez Longarela, I., Rsch, D., Rognone, L., Roseman, B., Rosu, I., Roy, S., Rudolf, N., Rush, S., Rzayev, K., Rzenik, A., Sanford, A., Sankaran, H., Sarkar, A., Sarno, L., Scaillet, O., Scharnowski, S., Schenk-Hopp, K. R., Schertler, A., Schneider, M., Schroeder, F., Schuerhoff, N., Schuster, P., Schwarz, M. A., Seasholes, M. S., Seeger, N., Shachar, O., Shkilko, A., Shui, J., Sikic, M., Simion, G., Smales, L. A., Sderlind, P., Sojli,

E., Sokolov, K., Spokeviciute, L., Stefanova, D., Subrahmanyam, M. G., Neusss, S., Szaszi, B., Talavera, O., Tang, Y., Taylor, N., Tham, W. W., Theissen, E., Thimme, J., Tonks, I., Tran, H., Trapin, L., Trolle, A. B., Valente, G., Van Ness, R. A., Vasquez, A., Verousis, T., Verwijmeren, P., Vilhelmsson, A., Vilkov, G., Vladimirov, V., Vogel, S., Voigt, S., Wagner, W., Walther, T., Weiss, P., van der Wel, M., Werner, I. M., Westerholm, P. J., Westheide, C., Wipplinger, E., Wolf, M., Wolff, C. C. P., Wolk, L., Wong, W.-K., Wrampelmeyer, J., Xia, S., Xiu, D., Xu, K., Xu, C., Yadav, P. K., Yage, J., Yan, C., Yang, A., Yoo, W., Yu, W., Yu, S., Yueshen, B. Z., Yuferova, D., Zamojski, M., Zareei, A., Zeisberger, S., Zhang, S., Zhang, X., Zhong, Z., Zhou, Z. I., Zhou, C., Zhu, X. S., Zoican, M., Zwinkels, R. C., Chen, J., Duevski, T., Gao, G., Gemayel, R., Gilder, D., Kuhle, P., Pagnotta, E., Pelli, M., Snksen, J., Zhang, L., Ilczuk, K., Bogoev, D., Qian, Y., Wika, H. C., Yu, Y., Zhao, L., Mi, M., Bao, L., Vaduva, A., Prokopczuk, M., Avetikian, A., and Wu, Z.-X. (2023). Non-standard errors. *Journal of Finance Forthcoming*.

Milgrom, P. and Stokey, N. (1982). Information, trade and common knowledge. *Journal of Economic Theory*, 26(1):17–27.

Offerman, T., Sonnemans, J., Van de Kuilen, G., and Wakker, P. P. (2009). A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *The Review of Economic Studies*, 76(4):1461–1489.

Okoli, C. and Pawlowski, S. D. (2004). The delphi method as a research tool: an example, design considerations and applications. *Information & Management*, 42(1):15–29.

Palley, A. B. and Satopää, V. A. (2023). Boosting the wisdom of crowds within a single judgment problem: Weighted averaging based on peer predictions. *Management Science*, forthcoming.

Peker, C. (2022). Extracting the collective wisdom in probabilistic judgments. *Theory and Decision*.

Popper, K. (1959). *The Logic of Scientific Discovery*. Routledge.

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.

Radanovic, G. and Faltings, B. (2013). A robust bayesian truth serum for non-binary signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27.

Radanovic, G. and Faltings, B. (2014). Incentives for truthful information elicitation of continuous signals. In *AAAI Conference on Artificial Intelligence*.

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):71–91.

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, 30(2):344–356.

Schlag, K. and Tremewan, J. (2021). Simple belief elicitation: An experimental evaluation. *Journal of Risk and Uncertainty*, 62:137–155.

Swank, J., Swank, O. H., and Visser, B. (2008). How committees of experts interact with the outside world: some theory, and evidence from the fomc. *Journal of the European Economic Association*, 6(2-3):478–486.

Tereick, B. (2020). Improving information aggregation through meta-cognition. Working paper.

Visser, B. and Swank, O. H. (2007). On committees of experts. *The Quarterly Journal of Economics*, 122(1):337–372.

Wilkening, T., Martinie, M., and Howe, P. D. (2022). Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems. *Management Science*, 68(1):487–508.

Witkowski, J. and Parkes, D. C. (2012). A robust bayesian truth serum for small populations. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

For Online Publication

A. Parameter values for each task and checking the incentive properties

In our experimental design, $P(S = Y) = P(S = N) = \frac{1}{2}$, and group size $n = 200$. Table 3 lists parameter values ω_N and ω_Y of the 30 tasks in our experiment. They contain all combinations where ω_N, $\omega_Y \in \{0.05, 0.1, ..., 0.45\}$, ω_N, $\omega_Y \in \{0.55, 0.6, ..., 0.95\}$, or $\omega_N \in \{0.05, 0.1, ..., 0.4\}$ and $\omega_Y \in \{0.6, 0.65, ..., 0.95\}$, and where the signal technology satisfies $\omega_Y - \omega_N \in \{0.2, 0.25, 0.3\}$ and $\bar{\omega}_1 - \bar{\omega}_0 > 0.04$.

We next show that the incentive properties established in Proposition 1 under an infinite group size are preserved under these parameter values when $n = 200$. We achieve this by first deriving the conditions for reservation prices if truth-telling equilibrium exists, and then showing that these reservation prices exist under our parameter values when $n = 200$.

Since the signals are drawn from the urn without replacement, the number of participants receiving signal 1 is fixed at $n\omega_S$ for $S = \{Y, N\}$. Suppose that all agents are truth-telling, i.e., $e_i = s_i$ for all i. Then the asset value can only be ω_S for $S = \{Y, N\}$. Let $\mathcal{E}_1(p)$ be the event that there exists an agent j such that $s_j = 1$ and $p_j \geq p$ and let $\mathcal{E}_0(p)$ be the corresponding event in which there exists an agent j such that $s_j = 0$ and $p_j \leq p$. Agent i after receiving signal 1 has the following expectation about the asset value given trade:

$$\begin{aligned} & E\left[v|\mathcal{E}_0(p_i), s_i = 1\right] \\ =& \omega_Y P(v = \omega_Y | s_i = 1, \mathcal{E}_0(p_i)) + \omega_N P(v = \omega_N | s_i = 1, \mathcal{E}_0(p_i)) \\ =& \omega_N + (\omega_Y - \omega_N) P(S = Y | s_i = 1, \mathcal{E}_0(p_i)), \end{aligned} \tag{11}$$

Using Bayes' rule, the conditional posterior can be further expressed as:

$$\begin{aligned} & P(S=Y|s_i=1,\mathcal{E}_0\left(p_i\right)) \\ =& \frac{\omega_Y\left(1-(1-p_i)^{n(1-\omega_Y)}\right)}{\omega_Y\left(1-(1-p_i)^{n(1-\omega_Y)}\right)+\omega_N\left(1-(1-p_i)^{n(1-\omega_N)}\right)}, \end{aligned} \tag{12}$$

where we used $P(S=Y)=P(S=N)=\frac{1}{2}$ and $P(s_i=1,\mathcal{E}_0\left(p_i\right)|S=Y)=P(s_i=1|S=Y)P(\mathcal{E}_0\left(p_i\right)|S=Y)$. Plugging this expression in Equation (11), we define for each task, a reservation price p_1^* for buyers, such that $E\left[v|s_i=1,\mathcal{E}_0\left(p_i\right)\right]>p_i$ when $p_i<p_1^*$, $E\left[v|s_i=1,\mathcal{E}_0\left(p_1^*\right)\right]=p_1^*$, and $E\left[v|s_i=1,\mathcal{E}_0\left(p_i\right)\right]<p_i$ when $p_i>p_1^*$. Similarly, we can also define a reservation price p_0^* for sellers. If p_1^* and p_0^* exist, and are such that $p_1^*>p_0^*$, then signal-1 agents will buy at price $p\leq p_1^*$ and signal-0 agents will sell at price $p\geq p_0^*$. These strategies constitute a truth-telling equilibrium.

We derive p_0^* and p_1^* for all the 30 tasks, which are shown in the last two columns of Table 3. Note that $\overline{\omega}_0$ and $\overline{\omega}_1$ defined in Equations (3)–(4) yields essentially the same values as p_0^* and p_1^* when one type of signal is not too rare, as in tasks 25–30. However, when one type of signal is rare, conditioning on the occurrence of trade is not negligible. For the parameters chosen in the experiment, a group size of 200 is still big enough for a trade interval to exist and thus preserving the incentive properties.

Table 3: Task parameter values

set	task	ω_N	ω_Y	$\overline{\omega}_0$	$\overline{\omega}_1$	p_0^*	p_1^*
1	1	0.05	0.25	0.14	0.22	0.20	0.22
2	2	0.05	0.30	0.16	0.26	0.20	0.26
1	3	0.05	0.35	0.17	0.31	0.20	0.31
2	4	0.10	0.30	0.19	0.25	0.21	0.25
1	5	0.10	0.35	0.20	0.29	0.22	0.29
2	6	0.10	0.40	0.22	0.34	0.23	0.34
1	7	0.15	0.35	0.24	0.29	0.25	0.29
2	8	0.15	0.40	0.25	0.33	0.26	0.33
1	9	0.15	0.45	0.27	0.38	0.27	0.37
2	10	0.20	0.40	0.29	0.33	0.30	0.33
1	11	0.20	0.45	0.30	0.37	0.31	0.37
2	12	0.25	0.45	0.33	0.38	0.34	0.38
2	13	0.75	0.95	0.78	0.86	0.78	0.80
1	14	0.70	0.95	0.74	0.84	0.74	0.80
2	15	0.65	0.95	0.69	0.83	0.69	0.80
1	16	0.70	0.90	0.75	0.81	0.75	0.79
2	17	0.65	0.90	0.71	0.80	0.71	0.78
1	18	0.60	0.90	0.66	0.78	0.66	0.77
2	19	0.65	0.85	0.71	0.76	0.71	0.75
1	20	0.60	0.85	0.67	0.75	0.67	0.74
2	21	0.55	0.85	0.63	0.73	0.63	0.73
1	22	0.60	0.80	0.67	0.71	0.67	0.70
2	23	0.55	0.80	0.63	0.70	0.63	0.69
1	24	0.55	0.75	0.62	0.67	0.62	0.66
1	25	0.30	0.60	0.41	0.50	0.41	0.50
2	26	0.35	0.60	0.45	0.51	0.45	0.51
1	27	0.35	0.65	0.46	0.55	0.46	0.54
2	28	0.40	0.60	0.48	0.52	0.48	0.52
1	29	0.40	0.65	0.49	0.55	0.49	0.55
2	30	0.40	0.70	0.50	0.59	0.50	0.59

B. Truth-telling at the individual level

Table 4 shows the proportion of participants with at least certain numbers of truth-telling in both SPA and FTM treatments.

Table 4: Proportion of participants with at least certain numbers of truth-telling

at least	SPA	FTM
1	100.0%	99.8%
2	100.0%	99.8%
3	100.0%	99.8%
4	99.6%	99.8%
5	99.3%	99.4%
6	98.5%	98.9%
7	97.8%	97.8%
8	96.0%	96.1%
9	92.8%	93.8%
10	90.5%	90.3%
11	86.7%	85.9%
12	81.9%	81.9%
13	77.6%	78.1%
14	71.3%	71.3%
15	58.5%	59.4%

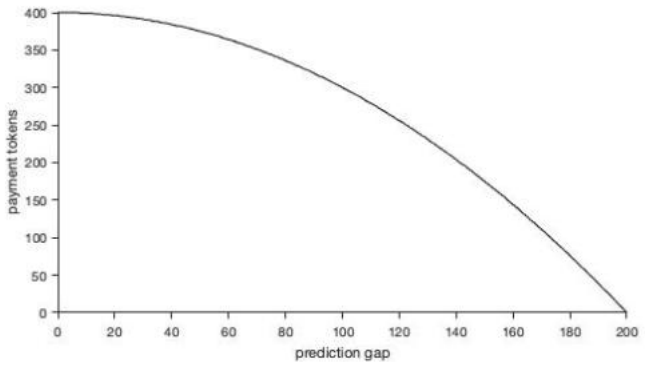
C. Experimental instructions and training rounds

We acquired IRB approval for our experiment design from Erasmus School of Economics, under the approval number ETH2122-0805, on July 7, 2022. The experiment and analysis plan were also preregistered (https://osf.io/cf8bk/) before the experiment was conducted on July 24, 2022.

We recruited participants from the experimental platform Prolific using the following pre-screening crteria: 1) fluent languages include English; 2) nationality is United States; 3) did not participate in our pilot experiments; 4) student status is yes. We estimated our experiment to last about 15 minutes, and the maximum allowed time for participants was 60 minutes. We informed participants beforehand in the description of the study that the fixed reward is £1.50, but they could also earn a bonus (up to £3) based on their answers. The experiment was conducted during July 24 to August 9, 2022.

All participants first watched the experimental instruction video (YouTube link) where the experimental setting of urns and balls was explained. Then they went through five rounds of training, first facing a task as displayed in 1–3 of the main text and then receiving feedback about how the payment was calculated. Figure 7 shows an example from the SPA treatment, and Figure 8 shows an example from the FTM treatment when the bet went through.

Your payment for this task depends on your prediction gap, which is the distance between your prediction and the actual number of participants who have guessed the same urn as you. **The less the prediction gap is, the higher your payment will be.**



Your payment for this task: 348 tokens

At the end of the experiment, we sum up your total payment tokens from all the 15 tasks and determine your bonus. Your bonus (in £) equals the total number of payment tokens divided by 2000. You could earn up to £3 in bonus.

Figure 7: An example of feedback in the training rounds in SPA.

The earnings of the bet (in tokens) is this number minus 110:
Earnings = 160 - 110 = 50

To make sure you don't lose money (if the earnings are negative), you will be endowed with 200 tokens, whether you take the bet or not.

You chose to take the bet.

Your bet goes through if someone took the opposite bet, which means, someone bet that less than 110 participants chose Urn Right (or in other words, more than 90 chose the other urn).

Your bet went through. Your payment for this task is 200 + earnings = 250 tokens.

If you had not taken the bet, then your payment would have been the endowment, 200 tokens.

At the end of the experiment, we sum up your total payment tokens from all the 15 tasks and determine your bonus. Your bonus (in £) equals the total number of payment tokens divided by 2000. You could earn up to £3 in bonus.

Note that the value 110 in the bet is randomly drawn.

Figure 8: An example of feedback in the training rounds in FTM.

Then each participant went through one set of 15 tasks, randomly chosen from the two sets of tasks with parameter combinations shown in Online Appendix A. Exact wording of the tasks is shown via screenshots in Figures 1–3 of the main text.

The last part of our experiment elicits some basic information about the participants (see Figure 9 for details). The export of the complete survey from Qualtrics is provided as supplementary material.

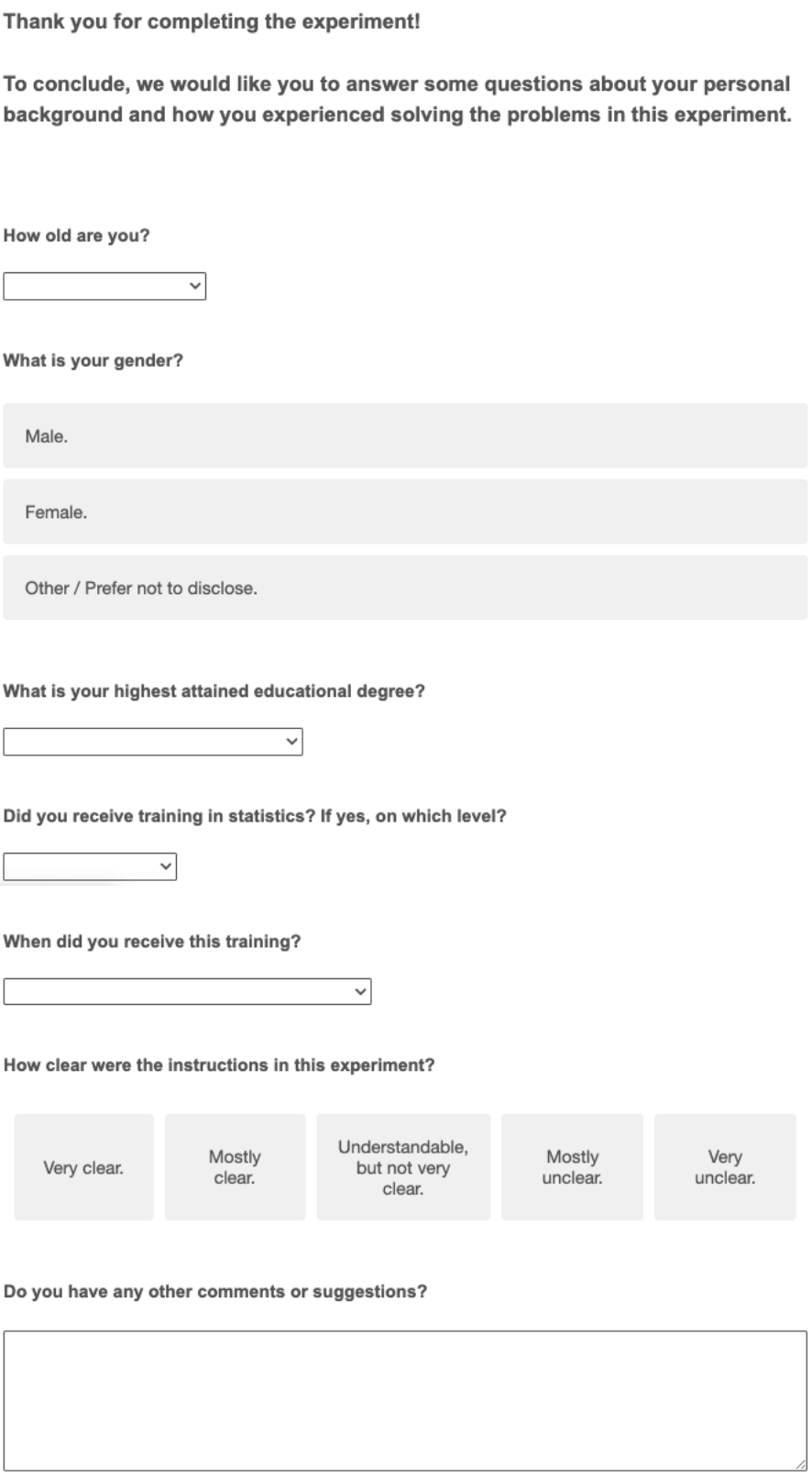


Figure 9: Exit survey.