Crowd Prediction Systems: Markets, Polls, and Elite Forecasters

Pavel Atanasov[a], Jens Witkowski[b], Barbara Mellers[c], Philip Tetlock[d]

[a] *Corresponding Author, Email: pavel@pytho.io, Pytho LLC, 866 President St, Brooklyn, NY, 11215, U.S.A.*
[b] *Frankfurt School of Finance & Management, Adickesallee 32–34, Frankfurt am Main, 60322, Germany*
[c] *University of Pennsylvania, 3730 Walnut Street, Philadelphia, 19104, PA, U.S.A.*
[d] *University of Pennsylvania, 3620 Locust Walk, Philadelphia, 19104, PA, U.S.A.*

Abstract

What systems should we use to elicit and aggregate judgmental forecasts? Who should be asked to make such forecasts? We address these questions by assessing two widely-used crowd prediction systems: prediction markets and prediction polls. Our main test compares a prediction market against team-based prediction polls, using data from a large, multi-year forecasting competition. Each of these two systems uses inputs from either a large, sub-elite or a small, elite crowd. We find that small, elite crowds outperform larger ones, whereas the two systems are statistically tied. In addition to this main research question, we examine two complementary questions. First, we compare two market structures, continuous double auction (CDA) markets and logarithmic market scoring rule (LMSR) markets, and find that the LMSR market produces more accurate forecasts than the CDA market, especially on low-activity questions. Second, given the importance of elite forecasters, we compare the talent-spotting properties of the two systems, and find that markets and polls are equally effective at identifying elite fore-

casters. Overall, the performance benefits of "superforecasting" hold across systems. Managers should move towards identifying and deploying small, select crowds to maximize forecasting performance.
Keywords: Forecasting, Judgment, Crowdsourcing, Aggregation, Markets

1. Introduction

Many forecasting processes necessarily rely on human predictive judgment. Crowd prediction systems, such as prediction markets, provide the infrastructure to elicit and combine the predictions from a group ("crowd") of forecasters. In contrast to purely data-driven approaches, crowd predictions are particularly important in settings with little historical data, such as in new product development (Cowgill and Zitzewitz 2015, Atanasov et al. 2023) and when predicting macro events, such as pandemics (Polgreen et al. 2007) or geopolitical developments (Tetlock and Gardner 2016).

Our examination focuses on a central practical question: if a manager seeks to maximize forecasting performance from a crowd of forecasters, what combination of crowd type and prediction system should she employ? Furthermore, we examine the case when both systems and crowds can evolve over time. More specifically, managers may choose to move away from a status quo system in favor of another, or select only a subset of individuals based on historical performance.

Our main research contribution lies in quantifying the impact of prediction system architecture and individual forecaster track record on aggregate performance. Previous research studied how the performance of CDA prediction markets and prediction polls compares when populated by sub-elite

forecasters (Atanasov et al. 2017), while published studies on elite forecasters have focused only on their individual performance (Mellers et al. 2015b). We are the first to compare the aggregate performance of small, elite forecaster crowds across two prediction systems: LMSR prediction markets and team prediction polls. Moreover, we compare the aggregate accuracy of elite forecaster crowds to larger, sub-elite crowds using the same prediction systems. The comparison of elite crowds is notable because such a study relies on the resource-intensive process of identifying elite forecasters: it involves engaging with thousands of forecasters reporting on hundreds of questions over multiple years. Studies involving fewer forecasters may need to set lower thresholds for elite status, while studies with fewer questions per season may identify top performers less reliably. This begs the question of whether forming elite forecaster crowds is worth the effort. Our results show that the benefits of employing elite crowds are large and robust across prediction polls and prediction markets, despite the threefold size advantage of sub-elite crowds. Moreover, the advantages of elite over sub-elite crowds are substantially larger than the differences between prediction markets and prediction polls.

In addition to this primary study, we also report on two additional studies, each of which complements the findings of the primary study along a different dimension. In the first, we provide an experimental evaluation of two popular types of prediction market architectures: continuous double auction (CDA) markets and logarithmic market scoring rules (LMSR) markets. To the best of our knowledge, we are the first to study these methods in a large, randomized experiment. Prior research reporting on CDA and LMSR

market performance did not compare the two designs directly but had separate sets of questions for each (Cowgill and Zitzewitz 2015). Using data from over 1300 forecasters and a total of 147 questions, we find that the LMSR market achieves higher accuracy than the CDA market. We find that the outperformance by the LMSR market appears particularly pronounced for questions that attracted few traders or soon after a question is posted, when only few traders had placed orders. Both of these correspond to thin markets and our analyses are hence in line with Hanson's (2003, 2007) main motivation for the design of the LMSR market architecture.

In the other complementary study, we examine the the relative effectiveness of different prediction systems in reliably identifying consistently accurate forecasters. Our data affords such an assessment as it features prediction markets and prediction polls running in parallel across three seasons. We show that prediction markets and prediction polls are equally effective at identifying elite forecasters for prediction polls. The practical consequence of this finding is that high earners in a prediction market are expected to continue to outperform when moved into an elite-crowd prediction poll.

1.1. Crowd Prediction Systems

The rationale for crowd prediction is based on two conceptual foundations. First, individual respondents have access to different *signals* (Silver 2012) about uncertain, future events. These signals vary in quality and aggregating them has the potential to integrate information that is dispersed among individuals. Second, individual judgments suffer from *noise* (Kahneman et al. 2021), i.e., "undesirable variability in judgments of the same problem." In the idealized case where noise is entirely due to judgmental er-

rors that are independent across participants, aggregation effectively reduces it as those noise terms "cancel out" (Surowiecki 2005). The benefits of aggregation are still present, albeit to a lesser extent, when errors in judgment are positively correlated across participants (Davis-Stober et al. 2014).

The two types of crowd prediction systems that have attracted the most attention from both researchers and practitioners are prediction markets (Wolfers and Zitzewitz 2004) and prediction polls (Winkler 1968), with the relative merits of these systems being subject to vigorous investigation (Graefe and Armstrong 2011, Atanasov et al. 2017, Reade and Williams 2019).

In a prediction market, traders buy and sell futures contracts that pay out if the corresponding event outcome occurs. The process of second guessing each other's bets yields market prices that correspond to probabilistic predictions. For example, in a binary-outcome prediction market for the U.S. presidential election, the contract corresponding to the Republican candidate may pay \$1 if the candidate wins, and \$0 otherwise. If the current market price is \$0.30, this would correspond to a probability of approximately 30% that the Republican candidate will win.

Prediction markets are conceptually based on the efficient market hypothesis (e.g., Malkiel and Fama 1970). The marginal trader hypothesis (Forsythe et al. 1992) further stipulates that a market with "a sufficient number" of traders tends to produce unbiased estimates even when most individuals are biased. Proponents argue that well-designed markets can generate accurate predictions for a wide range of future events, including macroeconomic data (Snowberg et al. 2013), sports results (Peeters 2018), election outcomes (Forsythe et al. 1992), and a variety of company-specific data (Spann

and Skiera 2003, Cowgill and Zitzewitz 2015).

Empirical tests demonstrate that markets can operate effectively even without real-money incentives (Servan-Schreiber et al. 2004) and with a limited number of forecasters (Healy et al. 2010). Researchers have noted that the predictive performance of markets depends more on market setup and less on the composition of the trader pool (Strijbis and Arnesen 2019).

Classic prediction markets are set up as continuous double auctions (CDA) and as such are generally expected to perform best in high-liquidity settings with many traders of varying skill (Forsythe et al. 1992). As an alternative to CDA markets, Hanson (2003, 2007) proposed logarithmic market scoring rules (LMSR). LMSR prediction markets use an automated market maker to address the challenges of (thin) markets with few traders. CDA markets have been studied more widely, especially since the creation of Iowa Prediction Markets (Berg and Rietz 2003), while the LMSR architecture is particularly well-adapted to settings with limited crowds. These two architectures have also been applied most widely in corporate settings (Cowgill and Zitzewitz 2015). Hence, they are the focus of the current research.

Prediction polls, also referred to as opinion pools or expert elicitation methods, and often used synonymously with forecasting tournaments, are an alternative crowd prediction system that relies on directly eliciting probability estimates from forecasters, providing proper-scoring feedback on their individual performance (Brier 1950, Murphy and Winkler 1987, Gneiting and Raftery 2007), and aggregating the individual estimates statistically (Satopää et al. 2014, Atanasov et al. 2017). For example, one forecaster in a prediction poll may submit a probability estimate of 40% while another submits an

estimate of only 10%. These forecasts would be combined by an aggregation algorithm, such as a weighted average or a more complex statistical model, and once the question is resolved, the forecasters would receive an accuracy score based on their individual forecasts and the outcome that materialized.

1.2. Large Crowds versus Small, Select Crowds

A separate line of research focuses on individual and group-level properties of accurate crowds. The group-level perspective on this question is that individual forecaster skill is less important than emerging properties of the crowd: large, diverse, and egalitarian crowds are expected to perform well. For example, the subtitle of Surowiecki's book (2005) emphasized the importance of crowd size: "Why the many are smarter than the few..." Similarly, Page (2007) stresses the value of diversity, and Davis-Stober et al. (2014) showed that crowds become wiser when crowd members make negatively correlated errors. Woolley et al. (2010) showed that groups' collective intelligence was more strongly correlated with the equality of contributions across group members than with their intelligence quotient (IQ) scores.

In contrast, the individual-level view is that accurate crowds are those made up of accurate individuals. Research from this perspective has shown that individual differences (e.g., fluid intelligence, cognitive styles, task engagement, and past performance) have a reliable association with individual performance on probabilistic prediction tasks (Mellers et al. 2015a, Tetlock and Gardner 2016). In contrast, apparent expertise, as assessed by education, professional experience, or eminence has surprisingly little relation to forecasting accuracy (Tetlock 2005). The strong form of this perspective is that crowd accuracy hinges primarily on "getting the right people on

the bus"—attracting, identifying, and retaining high performers. The main recommendation from this research is to employ small crowds of top forecasters, as assessed by their accuracy track records. In particular, Mannes et al. (2014) and Goldstein et al. (2014) showed that small, select crowds of forecasters with records of high achievement tend to outperform large, less selective crowds in a prediction poll setting. None of this research has thus far extended to a prediction market setting.

A variation on this elitist approach was taken by the Good Judgment Project (GJP), which selected the top 2% of forecasters each season, labeled them "superforecasters," (henceforth, "elite forecasters") and placed them in different teams competing in a prediction poll (Mellers et al. 2014). Team members could share information relevant to forecasting questions with one another and received team accuracy scores in addition to their individual accuracy scores (Mellers et al. 2015b, Tetlock and Gardner 2016). Elite forecasters outperformed sub-elite forecasters (the bottom 98% plus new, unproven individuals), and the approach played a key role in GJP's winning performance in the geopolitical forecasting tournaments sponsored by the Intelligence Advanced Research Project Activity (IARPA). Elite forecasters in team-based prediction polls were more accurate than professional intelligence analysts with access to classified information (Goldstein et al. 2016).

Notably, the previously published results on GJP elite forecasters focused exclusively on their individual performance in team-based prediction polls ("Superpolls"). But was the strong performance of Superpolls driven by its high-accuracy individuals or the team-based prediction polls architecture? To what extent did individual excellence result in superior aggregate

performance within the same prediction system? Would aggregate accuracy improve or worsen if crowds of elite forecasters worked individually in prediction markets ("Supermarkets") instead of Superpolls? We offer the first empirical assessment of these questions.

1.3. Research Questions

All of the research presented here focuses on the objective of identifying the combinations of prediction systems and forecasters that produce maximally accurate predictions. The investigation is organized around one main and two complementary research questions.

Main Research Question: Crowds and Prediction Systems

What is the impact on aggregate accuracy of crowd type (small, elite versus large, sub-elite) and prediction system (prediction markets versus polls)?

Persistent differences in accuracy among individual forecasters are well documented (Mellers et al. 2015a) and top performers also tend to make positive contributions to aggregate accuracy (Budescu and Chen 2015). We thus expect small, elite crowds to outperform larger, sub-elite crowds in prediction polls, and we examine if the number of forecasters is a limiting factor on aggregate accuracy in elite prediction polls. It is more difficult to predict if the potential benefits of employing small, elite crowds will carry over to prediction markets. It is plausible that limiting the number of traders may adversely impact activity or liquidity in prediction markets, offsetting the advantages of higher individual skill.

Regarding the comparison of prediction systems, prior results showed higher accuracy of prediction polls relative to CDA markets when both are

populated by sub-elite crowds (Atanasov et al. 2017). However, as discussed in Complementary Research Question 1, to the extent that the automated market maker of LMSR markets improves performance over that of CDA markets in small-crowd environments (Hanson 2003), differences in accuracy between prediction markets and prediction polls may be reduced. Finally, statistical aggregation algorithms with features such as temporal subsetting, accuracy weights, and extremization are known to improve prediction poll accuracy. We evaluate if the results of the comparison between small, elite and large, sub-elite crowds as well as between prediction markets and prediction polls depend on the the aggregation algorithm used in prediction polls.Our main finding is that small, elite crowds tend to produce consistently more accurate aggregate forecasts than non-elite crowds, whereas prediction markets and prediction polls are approximately tied in terms of accuracy.

Complementary Research Question 1: CDA versus LMSR Markets

Do LMSR prediction markets produce more accurate forecasts than CDA markets?

This question addresses the key argument that Hanson (2003) provided for the design of his logarithmic market scoring rules (LMSR) prediction market architecture, namely that CDA markets are expected to perform poorly in settings with insufficient trading activity. We discuss two potential moderators of differences in accuracy: the number of traders posting orders on the market for a given question and the timing within the question. Atanasov et al. (2017) showed that CDA markets underperform team-based prediction polls when question resolutions are months away but are approximately tied in accuracy in the last few weeks before question resolution. Markets are

complex adaptive systems (Markose 2005), with many factors contributing to aggregate performance. It is thus useful to quantify the extent to which theory-driven directional predictions bear out in experiments. Our empirical result is indeed consistent with the theoretical prediction, as it shows that LMSR markets outperform CDA markets, especially on low-activity forecasting questions. The results of this test also influence the study design employed for our Main Research Question; more specifically, the market structure employed in the comparison between prediction markets and prediction polls.

Complementary Research Question 2: Identifying Accurate Individuals

Which crowd prediction system is more effective at identifying accurate forecasters: prediction markets or prediction polls?

In addition to forecasts that support decision making, crowd prediction systems can also produce valuable information about the accuracy of individuals. For a performance assessment measure to be useful, it needs to produce similar rankings when the measure is collected again under similar conditions (i.e., it needs to be *reliable*). We assess the relative reliability of performance rankings in prediction markets and prediction polls. In the context of geopolitical forecasting tournaments, previous research has demonstrated that individual differences in prediction poll accuracy scores are reliable over time (Mellers et al. 2015a). Such accuracy measures provide useful inputs to weighted aggregation algorithms, which consistently outperform unweighted aggregation (e.g., Atanasov et al. 2017).

We know less about the association between prediction market earnings and forecaster accuracy as well as the reliability of market earnings and asso-

ciated rankings in prediction markets. Rothschild and colleagues (Rothschild and Sethi 2016, Schmitz and Rothschild 2019) show that traders exhibit reliable trading patterns; for example, only a minority engage in arbitrage trades, betting on both sides of a contract with minimal directional risk. The use of arbitrage strategies speaks to traders' engagement and relative sophistication in navigating the market environment but may or may not be associated with the traders' aptitude to generate predictive insight.

Practically speaking, a manager running a large prediction market would find little evidence in the literature on whether high-earning traders on the top of a prediction market leaderboard are reliably accurate forecasters, particularly skillful in executing trading strategies, or simply lucky.

We examine Complementary Research Question 2 at two levels: elite and sub-elite performance. First, at the elite level, we compare the accuracy of forecasters who have been identified as elite in a prediction poll with forecasters who have been identified as elite in a prediction market. In a later season, both groups provide probabilistic forecasts in a Brier-scored prediction poll. Since the trading skills required to excel in prediction markets may not perfectly align with the probability estimation skills needed to attain elite status in prediction polls (i.e., skill transfer between the prediction market trading environment and the prediction poll setting may be less than perfect), we expect that elite traders identified in prediction markets will underperform elite forecasters identified in prediction polls. Inconsistent with this expectation, our results show that elite forecasters perform at similar levels, independently of whether they qualified through their top performance in prediction markets or polls.

Second, in our data, forecasters who do not reach elite levels (do not place in the top 2%) tend to remain in the same system across seasons. This allows us to compare the reliability of rankings across subsequent seasons of the tournament for both prediction markets and prediction polls. In contrast to Brier scores in prediction polls, where scores of individual questions are limited to the range between 0 and 2, prediction market earnings may depend more on the outcome of a single question, and thus luck may be more important—a single large bet can account for a large proportion of a trader's gains or losses. Therefore, we expect the cross-season reliability of market earnings to be lower than the cross-season reliability of Brier scores from prediction polls.[1] Our empirical results are consistent with this expectation.

2. Methods

All data were collected in the IARPA Aggregative Contingent Estimation (ACE) tournament (2011–2015), a forecasting tournament that consisted of 4 forecasting seasons, each lasting approximately 9 months (Good Judgment Project 2016). Our Primary Research Question focuses on aggregate performance in Season 4. We compare the performance of four separate groups using a two-factor design: forecaster accuracy (elite versus sub-elite) and crowd prediction system (LMSR prediction market versus prediction poll). Data from Seasons 1, 2, and 3 were used to identify elite forecasters and to estimate optimal parameters for the prediction poll aggregation algorithms.

[1] While we consider the reliability of performance rankings to be an important property of the system, we note that performance ranks do not equate to forecasters' individual contributions to overall accuracy.

Complementary Research Question 1 uses data from CDA and LMSR markets from Season 3 of the tournament. Complementary Research Question 2 focuses on individual performance across markets and polls. At the elite level, we examine the accuracy of forecasters in Season 3 in Superpolls, comparing those who attained elite status through their performance in Season 1 and 2 through prediction markets versus those who qualified through prediction polls. Reliability of sub-elite performance is assessed using data from Seasons 2, 3, and 4. In any one season, a forecaster participated in only one prediction system.

2.1. Crowd Prediction Systems

Prediction Markets

Prediction markets run by GJP used play-money contracts valued between $0 and $100. When a question resolved, the price was set to $100 if the event occurred, and $0 otherwise. Each forecaster (trader) was provided with an initial endowment of $10,000. Leaderboards featured the top 50 forecasters based on their total balance. The aggregate probability forecast on a given question and day was the last price as of midnight Pacific Time.

Prediction markets in Seasons 2 and 3 were continuous double auction (CDA) markets in which forecasters traded with one another by placing bids and asks on the order book. Both price history and order book, which displayed the six highest bids and the six lowest asks, were public information. Logarithmic Market Scoring Rule (LMSR) markets (Hanson 2003, 2007) were employed in Season 3 (parallel to CDA markets) and Season 4 using software from Inkling Markets (now operating as Cultivate Labs). In these LMSR markets, forecasters traded with an automated market maker that was con-

stantly available to quote prices based on the current market price and the number of to-be-traded shares. Below is Pennock's (2006) implementation of the price function for a binary (yes/no) question:

$$\text{price}_{yes} = \frac{e^{q_{yes}/b}}{e^{q_{yes}/b} + e^{q_{no}/b}} \cdot \$100 \tag{1}$$

The current price of a contract for the "yes" outcome (price_{yes}) increases with the quantity of shares traded on that outcome (q_{yes}) and decreases with the quantity traded on the "no" outcome (q_{no}). The liquidity parameter b determines how prices respond to trading activity, with higher values corresponding to more liquidity and hence less price movement for a given quantity of traded shares. Based on Inkling's prior experience, the liquidity parameter was fixed at 250 and held constant across all questions in the tournament. For example, at $b = 250$, buying 100 shares in a binary market that is newly initiated at a price of \$50 for each outcome would move the price from \$50.00 to \$59.87, for an average price per share of \$55.02. Price history was publicly available to market participants, who could view their portfolio holdings, including profits and losses per question as well as the play money balance available for trading.

Prediction Polls

Forecasters provided probability forecasts in two variants of prediction polls: independent and team-based. The accuracy metric was the Brier score (Brier 1950), which varies from 0 (best) to 2 (worst). For questions with ordered response categories, we used the ordered scoring variation of the Brier score (Jose et al. 2009). Scores of individual forecasters were based on the average daily Brier score, which averaged scores across all days on

which the question was open for forecasting. If a forecaster did not update an estimate on a given day, her most recent forecast was carried over for scoring purposes. Scores for days up to the first forecast were imputed as that question's mean Brier score of all active forecasters in the same condition. If a forecaster did not report on a given question, the score for that question was imputed for the entire question duration. Brier scores based on individual questions were then averaged over questions and the 50 forecasters with the lowest (best) Brier scores were featured on a leaderboard. Forecasters submitted individual forecasts in both independent and team-based prediction polls—team consensus was not required. Team-based prediction polls differed from independent polls in that team members could communicate with one another. For each forecasting question, the team-level Brier score was the median of all team members' Brier scores. The leaderboards displayed both individual and team-level Brier scores as well as the corresponding rankings.

To combine individual estimates, we used the weighted mean algorithm described by Atanasov et al. (2017), with two additional weighting features that were based on a forecaster's psychometric test score as well as on the time she spent on the platform. A weighted logit algorithm (Satopää et al. 2014) was used in sensitivity analyses. The difference between the two is that the weighted mean algorithm averages forecasts in the original probability space, whereas the weighted logit algorithm first transforms forecasts into log-odds (logit), then averages them, and then converts them back to probabilities. Both algorithms feature (1) temporal subsetting, (2) differential forecaster weights, and (3) extremization. Temporal subsetting ensured that data was timely by including only forecasts from days containing the most

recent $k\%$ forecasts for a given question. Weighting increased the influence of forecasters with a track record of high accuracy as measured by z-score-transformed Brier scores, high forecast-updating frequency, more time spent on the platform, higher scores on psychometric measures of intelligence and political knowledge (Mellers et al. 2015b). Finally, extremization was applied to aggregate forecasts using the formula

$$\hat{p} = \frac{\bar{p}^a}{\bar{p}^a + (1 - \bar{p})^a}, \tag{2}$$

where $\hat{p}$ is the extremized probability estimate, $\bar{p}$ is the raw aggregate probability estimate, and a is the recalibration parameter. Note that $a = 1$ denotes the identity transformation. For $a > 1$, values are extremized (i.e., pushed away from 0.5 towards 0 or 1), and for $a < 1$, aggregate forecasts are made less extreme (i.e., pushed towards 0.5). This function was also applied to prediction market prices to assess if markets exhibit the well-known favorite-long-shot bias (Page and Clemen 2013).

The parameters for the aggregation algorithms were optimized at the start of a new season based on data from all previous seasons using elastic net regularization (Zou and Hastie 2005) and in the following order: temporal subsetting, forecaster weights, and extremization. The objective was to minimize aggregate forecast error. In the present work, analyses for our Main Research Question rely on aggregation in prediction polls. For the relevant Season 4, the estimated values using data from Seasons 1–3 for temporal subsetting resulted in using the $k = 20\%$ most recent forecasts for sub-elite team-based prediction polls, and the most recent $k = 53\%$ and $k = 73\%$ of forecasts for Superpolls logit and mean, respectively. Accuracy parameter settings were such that the most accurate forecaster received approximately

16 and 13 times the weight of the median-accuracy forecaster at the time, in sub-elite and elite teams, respectively. Other differential forecaster weights were of secondary importance. Extremization parameters were set at 1.5 and 1.32 for sub-elite teams and elite teams, respectively.

2.2. Participants

GJP forecasters were recruited from email lists, professional societies, research institutes, alumni associations, and by word-of-mouth. They were required to hold a bachelor's degree or higher. Before entering the tournament, they completed psychometric and political knowledge tests (lasting approximately two hours) as well as online training modules (lasting approximately one hour). Forecasters were mostly male (80%+) with a mean age of 36.

Financial incentives were provided for active participants, based on rules communicated at the start of each season. Every forecaster who made at least 25 forecasts received a $150 gift certificate in Season 1 and a $250 certificate Seasons 2, 3, and 4. Each returning forecaster in Season 2, 3, and 4 received an additional $100 gift certificate. Forecasters who placed in the top 2% of their condition's leaderboard were invited to become "superforecasters" (elite forecasters) in the subsequent season. All superforecasters were invited to travel-expenses-paid in-person workshops that took place at university campuses after Seasons 2, 3, and 4.

Table 1 displays all relevant study conditions across the four seasons of the forecasting tournament. For each of the research questions, we used all available data for the relevant conditions.

Our *Main Research Question* focuses on elite forecasters and sub-elite

Table 1: Good Judgment Project conditions across all 4 seasons.

	Season			
Condition	1	2	3	4
Independent Polls	✓	✓	✓	✓
Team Polls	✓	✓	✓	✓
Superpolls (Team Polls)		✓	✓	✓
CDA Markets		✓	✓	
LMSR Markets			✓	✓
Supermarkets (LMSR Markets)				✓

Note: Only Superpolls and Supermarkets conditions are populated by elite forecasters.

Table 2: Study design for Main Research Question; n denotes the number of forecasters in each cell. All data are from Season 4.

	Forecaster Type	
Prediction System	Elite	Sub-Elite
Prediction Markets	Supermarkets	Sub-Elite Markets
	$n = 122$	$n = 404$
Team Prediction Polls	Superpolls	Sub-Elite Polls
	$n = 139$	$n = 430$

forecasters in Season 4, the only season in which a sufficient number of elite forecasters were available to afford allocation across two conditions: team prediction polls (Superpolls) and LMSR markets (Supermarkets). Elite forecasters were identified in Seasons 1, 2, and 3 of the IARPA tournament based on their season-end performance. To qualify, a participant had to rank in the top 2% of their condition in a season. The ranking was based on the Brier score in prediction polls and end-of-season earnings in prediction markets. Once qualified, elite forecasters retained this status unless they dropped out or asked to rejoin the sub-elite crowd. Prediction markets with elite forecasters (Supermarkets) consisted of $n = 122$ traders working independently in an LMSR prediction market (also see Table 2). Team prediction polls with elite forecasters (Superpolls) consisted of 10 teams with 12 to 16 forecasters each, totaling $n = 139$ forecasters. For sub-elite forecasters, $n = 404$ were assigned to work independently in an LMSR prediction market and $n = 430$ were assigned to team prediction polls. Thus, for both prediction systems, sub-elite crowds were over three times larger than elite crowds.

Assignment of elite forecasters to Season 4 conditions was not random (also see Table 3). Forecasters who were active elite forecasters in Season 3 (i.e., attained elite status in Season 1 or 2) self-selected into Superpolls or Supermarkets, with the majority selecting into Superpolls (85/107). Forecasters who attained elite status based on Season 3 performance were assigned to Superpolls or Supermarkets mostly based on their Season 3 conditions. Five other elite forecasters who had not participated in Season 3 but had attained elite status previously were assigned to Superpolls.

Sub-elite forecasters were randomly assigned to conditions (also see Ta-

Table 3: Elite forecaster transition from Season 3 to Season 4.

	Season 4 Condition		
Season 3 Condition	Superpolls	Supermarkets	Total
Superpolls (self-selection)	85	22	107
Team Polls	45	2	47
Independent Polls	4	16	20
Prediction Markets	0	82	82
Other	5	0	5
Total	139	122	261

Table 4: Sub-Elite forecaster transition from Season 3 to Season 4.

	Season 4 Condition		
Season 3 Condition	Team Polls	LMSR Markets	Total
Team Polls	127	0	127
Independent Polls	0	28	82
Prediction Markets	34	130	164
None	243	272	515
Total	404	430	834

Note: All conditions are populated by sub-elite forecasters.

ble 4). The majority (62%) of Season 4 sub-elite forecasters joined the project in Season 4 and were newly assigned. The second largest group (31%) consisted of participants who returned from Season 3 and continued to participate in the prediction system (prediction market or prediction poll) to which they had been randomly assigned in Season 3. Finally, a small group (7%) were assigned to switch from markets to polls or from polls to markets between Seasons 3 and 4.

Complementary Research Question 1 focuses on the comparison of CDA and LMSR prediction markets in Season 3, the only season with both CDA and LMSR markets running side by side. Forecasters were randomly assigned to either the CDA ($n = 664$ forecasters) or LMSR ($n = 679$ forecasters) prediction market. Forecasters had the option of asking to have their data removed but few did. Mid-way through the season, additional forecasters were added, bringing the total to 750 assigned forecasters per market. Randomization was stratified by returnee status, balancing the number of new and experienced forecasters across the two markets. Because the CDA versus LMSR test took place in Season 3, while the LMSR versus prediction polls test took place in Season 4, we discuss these separately, rather than as parts of a single comparison.

Complementary Research Question 2 uses data from Seasons 2, 3, and 4 to assess the test-retest reliability of performance rankings among sub-elite forecasters. These were the three seasons in which prediction markets and prediction polls ran in parallel. We include all forecasters who competed in the same system (prediction markets or prediction polls) across two consecutive tournament seasons. For example, a forecaster would be included if

they competed in a prediction market both in Season 2 and Season 3 or if she competed in prediction polls both in Season 3 and Season 4. Forecasters who switched systems, attained elite status, or dropped out across seasons are not included. The analysis includes sub-elite forecasters competing in independent or team-based prediction polls in Seasons 2–3 ($n = 412$) and Seasons 3–4 ($n = 244$) as well as prediction market traders in Seasons 2–3 ($n = 237$) and Seasons 3–4 ($n = 508$). The differences in number of participants across conditions (systems) are a function of experimental assignment, rather than forecaster self-selection into conditions.

At the elite level, we use Season 3 data with $n = 126$ elite forecasters working in Superpolls. This was the only season in which all elite forecasters competed side by side in one condition, and where their prior season rankings were available from both prediction markets and prediction polls. The plurality had worked in Superpolls in Season 2 ($n = 49$). Newly qualified elite forecasters from Season 2 had worked in team-based prediction polls ($n = 26$), independent prediction polls ($n = 25$), prediction markets ($n = 21$), and other conditions ($n = 5$).

2.3. Forecasting Questions and Scoring

Seasons 1, 2, 3, and 4 featured 85, 114, 147, and 136 resolved questions, respectively. Forecasting questions were released throughout the forecasting season in batches of 1–10 questions at a time. Across all seasons, median question duration was 82 days (interquartile range: 40 to 153). Forecasters were encouraged to update their estimates as often as they wished until questions were resolved. Probability forecasts in prediction polls were aggregated and compared to market prices at the same time each day. Aggregate

forecasts were scored using the average daily Brier score—the same rule as for individual prediction poll forecasters. We perform sensitivity analyses in which logarithmic scores are used in place of Brier scores to assess aggregate accuracy. Logarithmic scores range from -∞ (worst) to 0 (best possible accuracy).

3. Results

3.1. Main Research Question: Individual Forecaster Accuracy versus Prediction System

To determine the impact of crowd type and prediction system on overall accuracy, we compare the performance of four separate groups using a two-factor design: individual forecaster accuracy (small, elite versus large, sub-elite) and prediction system (LMSR prediction market versus prediction poll). Within polls, we vary the sophistication of the aggregation algorithms between simple unweighted linear opinion pools ("ULinOP"), which corresponds to the simple average of all forecasters' most recent forecasts on a question, and the more complex, weighted algorithm ("full aggregation") described in Section 2.1. Inferential tests were based on mixed-effects models with random intercepts for forecasting questions, as implemented in the R nlme package.

Figure 1 summarizes the main results: crowd type and prediction system with different prediction poll aggregations. The largest and most notable accuracy difference is that between small, elite and large, sub-elite crowds. Team prediction polls with full aggregation show small advantages over prediction markets, and more sophisticated aggregation algorithms (full versus

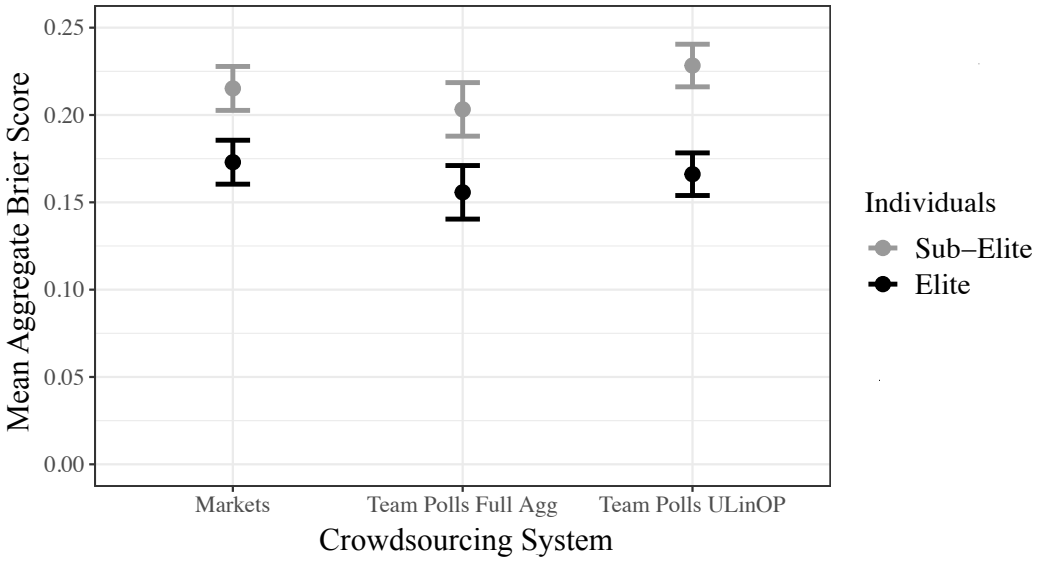

Figure 1: Aggregate accuracy on 136 questions in Season 4 by forecaster accuracy (sub-elite versus elite) and prediction system (prediction markets and team prediction polls). Lower scores denote better accuracy. Error bars denote one standard error of the difference in scores for each pair.

ULinOP) also yield small advantages within prediction polls.

We first examine accuracy differences among the four forecaster groups, and later return to the comparison of aggregation algorithms. Brier score descriptive statistics by individual accuracy level and prediction system are shown in Table 5. A sensitivity analysis with logarithmic scores is shown in Table 6 and largely replicates the Brier score pattern. Table 7 displays the results of the mixed models, the primary inferential tests addressing our Main Research Question.

In both prediction markets and team-based prediction polls, small elite crowds outperformed larger, sub-elite ones. The accuracy advantage of small, elite crowds amounted to approximately 0.05 on the Brier score scale ($b =$

Table 5: Aggregate-level Brier Scores (BS) for 136 questions in Season 4 by forecaster accuracy level (Elite versus Sub-elite) and prediction system (Prediction Polls versus Prediction Markets).

	Sub-Elite		Elite		% Elite	Cohen's d
	Mean	(SD)	Mean	(SD)	BS Adv.	
Pred. Markets	0.215	(0.280)	0.173	(0.242)	20%	0.16
Pred. Polls	0.203	(0.330)	0.156	(0.305)	23%	0.15
% Polls BS Adv.	6%		10%			
Cohen's d	0.04		0.06			

Note: Lower Brier scores denote better accuracy. Mean Brier score reduction and Cohen's d values are positive if the Brier score in the second row/column is lower than in the first row/column. See Figure 1.

Table 6: Aggregate-level Logarithmic Scores (LS) for 136 questions in Season 4 by forecaster accuracy level (Elite versus Sub-elite) and prediction system (Prediction Polls versus Prediction Markets).

	Sub-Elite		Elite		% Elite	Cohen's d
	Mean	(SD)	Mean	(SD)	BS Adv.	
Pred. Markets	-0.350	(0.358)	-0.300	(0.322)	14%	0.14
Pred. Polls	-0.328	(0.470)	-0.270	(0.513)	18%	0.12
% Polls LS Adv.	6%		10%			
Cohen's d	0.05		0.09			

Note: Higher (less negative) scores denote better accuracy. LS improvement and Cohen's d values are positive if the logarithmic score in the second row/column is higher than in the first row/column.

Table 7: Mixed-effects models: aggregate Brier score by crowd type and prediction system, based on 136 questions in Season 4.

DV: Season 4 Aggregate Brier Score	A. Main Effects	B. Interaction
Intercept	0.217 (0.025)	0.215 (0.025)
Forecaster Accuracy		
Sub-Elite (Reference)		
Elite	-0.045 (0.010) **	-0.042 (0.014) **
Prediction System		
Prediction Markets (Reference)		
Prediction Polls	-0.015 (0.010)	-0.012 (0.014)
Interaction: Elite × Polls		-0.005 (0.019)
AIC	-369.10	-361.13

Note: * $p < .05$, ** $p < .01$. Lower values denote better performance.

-0.045, $se = 0.010$, $t = -4.62$, $p < .001$), corresponding to an accuracy improvement of 21%. Prediction markets and prediction polls did not differ significantly in accuracy ($b = -0.015$, $se = 0.010$, $t = 1.51$, $p = .13$). See Figure 1. There was no significant interaction between individual forecaster accuracy type and prediction system. (See Table 7, Column B).

Addressing our Main Research Question, the effect of individual forecaster track record on aggregate accuracy was large and significant, while the choice of prediction system did not correspond to significant differences in accuracy.

Within prediction polls, we also examined how the effects of individual forecaster accuracy compare to those of using more or less sophisticated aggregation algorithms. In particular, we compared the full algorithm featuring temporal subsetting, forecaster weights, and extremization (as explained in Section 2.1) with the simple unweighted linear opinion pool (ULinOP). The improvements in accuracy from using the full algorithm versus ULinOP were smaller than the improvements in accuracy from employing elite versus sub-elite forecasters. In fact, the simple ULinOP of small, elite crowds outperformed the full aggregation algorithm of large, sub-elite crowds, yielding 15% lower Brier scores.[2]

[2]There were no differences between the two fully optimized aggregation algorithms, mean and logit. Moreover, neither Superpoll mean (Brier score Mean= 0.156) nor Superpoll logit algorithms (Brier score Mean= 0.157) significantly outperformed Supermarkets in accuracy at $\alpha = 0.05$. These results are based on mixed models of the type shown in Table 7 but using only data from Superpolls and Supermarkets. We focus on the weighted mean aggregation for simplicity but all results also hold for the logit algorithm.

Table 8: Brier score decomposition based on 136 questions in Season 4 by individual forecaster accuracy (elite versus sub-elite) and prediction system (prediction polls versus prediction markets). Lower calibration error, higher discrimination and lower simple Brier Scores denote better performance.

	Calibration	Discrimination	Uncertainty	Simple BS
Superpolls	0.009	0.44	0.61	0.18
Sub-elite Polls	0.008	0.37	0.61	0.25
Supermarkets	0.021	0.43	0.61	0.20
Sub-elite Markets	0.011	0.38	0.61	0.24

Note: Smaller calibration error and larger discrimination values denote better performance. Simple Brier scores do not account for ordered outcomes and are thus higher than ordered Brier scores in Table 7.

Calibration and Discrimination

Brier score decomposition (Murphy and Winkler 1987) analyses show that elite and sub-elite forecaster crowds registered similar calibration errors in both prediction markets and prediction polls. The differences in accuracy were entirely accounted for by elite crowds' superior discrimination scores. See Table 8.

Crowd Size Sensitivity

A key concern regarding elite crowds is the limited number of forecasters. In our data, the cutoff was set so that only the top 2% of forecasters were invited to become elite forecasters. This cutoff was chosen in light of the unique tournament constraints so it is not meant to be universal.

Since elite crowds were treated differently (they worked together), we cannot simulate performance in counterfactual scenarios with less restrictive cutoffs (e.g., 5%). Furthermore, prediction market interactions are cumulative since traders react to price changes, so we cannot simulate how traders would have reacted to a smaller or larger number of traders. We can, however, simulate how performance is affected by the number of Superpoll teams included in the aggregation because teams were disincentivized from sharing information with other teams and independence across teams is a reasonable assumption.

To make this analysis more applicable across contexts, we find it useful to consider the ratio between the number of forecasters and the number of questions in a given season. In Season 4, there were 139 Superpoll forecasters and 136 resolved forecasting questions, for a forecaster-to-question ratio of approximately 1:1. We reran the aggregations with subsets of the 10 Superpoll teams. We performed aggregation and scoring for combinations of 2 to 9 teams, corresponding to approximate forecaster-to-question ratios between 1:5 and 9:10. For each number of teams, we produced 10 scoring iterations, sampling teams without replacement in each iteration. For aggregation, the simple ULinOP algorithm was used for all subsets because the full algorithms' parameters were optimized only for the full sample.

Figure 2 displays the results across 10 iterations. The Superpoll ULinOP including all teams achieved a Brier score of 0.166. When teams were subsampled so that the forecaster-to-question ratios were reduced, mean Brier scores increased to a maximum of 0.176, corresponding to an accuracy reduction of 6%. Aggregate accuracy for elite crowds remained significantly superior to that of over 300 ULinOP-aggregated sub-elite crowd forecasters

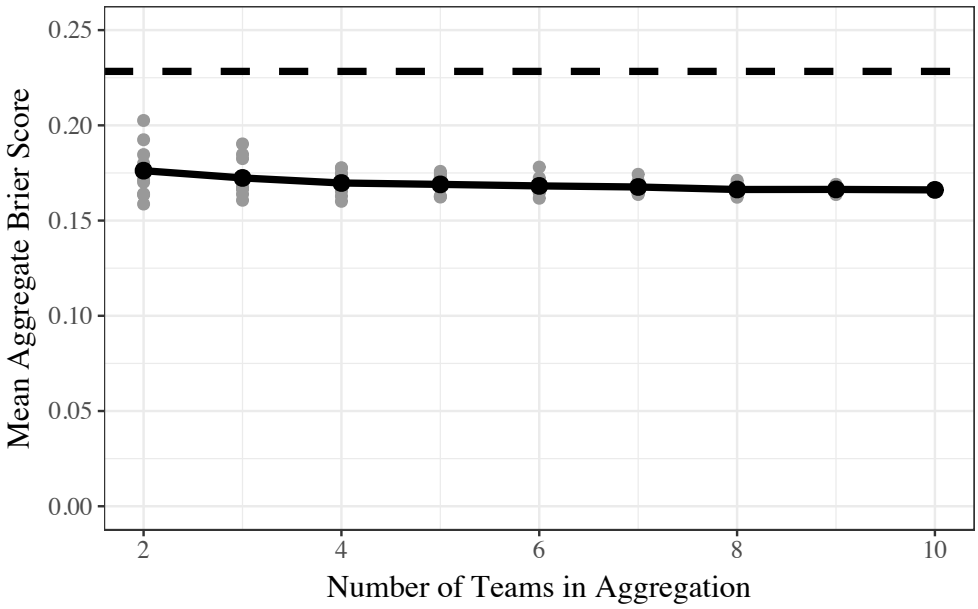

Figure 2: Aggregate accuracy on 136 questions in Season 4 by number of Superpoll teams included in the aggregation. Gray dots represent one scoring sample, while the black line and black dots show the mean scores across aggregations. The dashed line represents the accuracy of the ULinOP of the full-sample sub-elite team poll.

(dashed line, mean Brier score of 0.228) even when only two teams (24–28 elite forecasters) remained in the crowd. Moreover, any one of the ten subsets of two Superpoll teams we tested would have outperformed even the full-sample, fully optimized aggregation of sub-elite team polls (not shown).

Extremized Supermarket Prices

Superpoll algorithms had a potential advantage over the Supermarket: algorithms included aggregate-level extremization whereas market prices were not extremized. If extremization were helpful in improving Supermarket accuracy, that would suggest that Supermarkets' (non-significant) underperformance versus Superpolls is partly due to easily correctable miscalibration and not a fundamental deficit of predictive insights generated by Supermarket traders.[3] We thus examined different degrees of market price extremization. First, we applied the same extremization level used in Superpoll mean algorithms ($a = 1.32$) to Supermarkets. Second, we backtested Season 3 LMSR market prices and found that the same level of extremization ($a = 1.32$) minimized Brier scores. Finally, we applied the optimal-in-hindsight extremization level for Season 4 ($a = 1.75$). Note that this is not a realistic estimate of real-world performance but was included to assess the maximum potential benefit of extremization.

Table 9 displays the results. After applying the $a = 1.32$ extremization,

[3]Another approach for dealing with underconfident forecasts in LMSR markets is to reduce the liquidity parameter, which leads to larger prices movements for a given order size. We cannot simulate price movements with a different liquidity parameter ex post but these extremization analyses provide a different way to estimate the impact of underconfidence in the LMSR Supermarket.

Table 9: Season 4 Performance for different levels of extremization in Supermarkets.

Extremization Parameter Source	a	Mean BS	SD BS
No Extremization, Default	1.00	0.173	0.242
Season 3 Superpoll & LMSR Prediction Market	1.32	0.161	0.273
Season 4 Hindsight Optimization	1.75	0.158	0.301

the overall Supermarket Brier score improved from 0.173 to 0.161. Applying the hindsight-optimized extremization ($a = 1.75$) would have reduced Supermarket Brier scores further to 0.158, approximately equivalent to the 0.156 for the extremized, full-algorithm Superpolls aggregation.

3.2. Complementary Research Question 1: CDA versus LMSR Markets

The two markets attracted similar activity levels. On the CDA market, the median number of traders per question was 80 ($M = 87.7$, $SD = 46.4$), while on the LMSR market the median was 74 ($M = 84.9$, $SD = 44.4$). The two types of markets also attracted similar order volumes with an average of approximately 300 orders per question for both the CDA ($M = 315.3$, $SD = 350.2$) and the LMSR market ($M = 298.8$, $SD = 285.7$).

Across the 147 resolved questions in Season 3, the LMSR market earned lower (i.e., better) Brier scores ($M = 0.211$, $SD = 0.280$) than the CDA market ($M = 0.245$, $SD = 0.327$). The differences in Brier scores were significant in a paired t-test ($t(146) = 2.28$, $p = 0.024$). This is equivalent to a 14% Brier score reduction for the LMSR market relative to the CDA market (Cohen's $d = 0.12$). These results are also consistent with those obtained

Table 10: Mixed-effects models: Aggregate Brier scores by prediction market type (CDA versus LMSR) and timing within question in Season 3. Timing main and interaction effects are estimated for absolute number of days and proportion of time within a question.

DV:	A.	B.	C.
Agg. Brier Score	Main Effect	Abs. Time	Prop. Time
Intercept	0.231 (0.024)	0.166 (0.029)	0.185 (0.027)
CDA (Reference)			
LMSR	-0.026 (0.002)**	-0.022 (0.003)**	-0.009 (0.004)*
Days to Res. (x100)		0.204 (0.050)**	
LMSR × Abs. Time		-0.005 (0.003)	
Prop. Time to Res.			0.094 (0.024)**
LMSR × Prop. Time			-0.034 (0.007)
AIC	-16719.7	-16701.5	-16735.7

Note: * $p < .05$, ** $p < .01$. Lower values denote better performance.

by a regression specification that uses Brier scores for each day within a question, rather than only one score per question, utilizing a mixed-effects model with random question intercepts. (See Table 10, Column A.)

What were the main sources of the CDA market's relative underperformance? The experiment was not designed to specifically address this question, but we conducted exploratory analyses, providing some indication. The larger standard deviations of Brier scores for the CDA market point to the

possibility that the CDA market's underperformance was driven by a small number of high-Brier-score questions. Our expectation—as indicated by the literature (Hanson 2003, 2007)—was that CDA markets may underperform in thin-market settings. In the tournament, traders self-selected into questions, so we do not have the benefit of random assignment of forecasters to questions. However, we can still examine if the CDA market underperformed on questions that attracted fewer traders. Questions on which CDA markets attracted a larger number of traders also tended to attract a larger number of traders in the LMSR market ($r = 0.88$). Hence, we use the average number of traders across the two markets as our measure of activity on a question. Results are similar when using the number of CDA traders or LMSR traders instead.

Figure 3 shows the results of this exploratory analysis. The outcome measure is the difference in Brier scores between the LMSR and the CDA markets, where positive values denote better Brier scores for the LMSR market relative to the CDA market (and vice versa). On questions attracting larger numbers of traders in the CDA market, Brier score differences consistently clustered around zero. On questions with few traders, however, Brier score differences were larger and more variable. Notably, the CDA market registered its worst relative performances on low-activity questions attracting fewer than 100 traders. A simple correlational analysis reveals that Brier score differences are positively correlated with the number of CDA traders (Pearson's $r = 0.15, p = .062$) and the mean number of traders across the two markets (Pearson's $r = 0.17, p = .035$), denoting that CDA market underperformance tended to observed on questions with fewer CDA traders. While

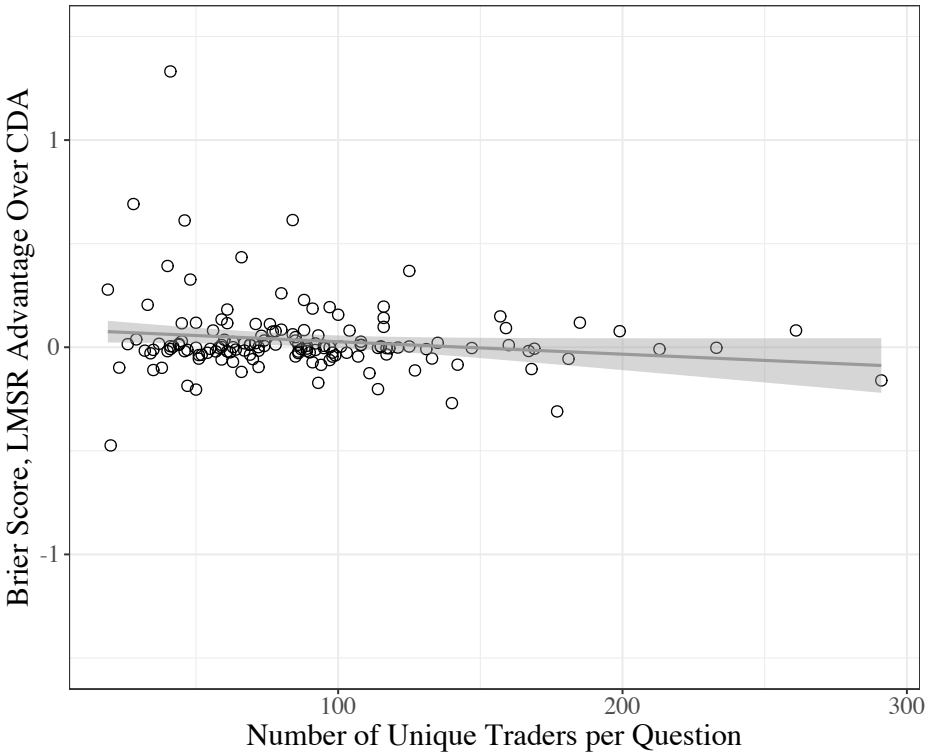

Figure 3: Differences in Brier scores between the LMSR and CDA market, one point per question, plotted against the mean number of traders posting orders on the CDA market. Positive score differences denote LMSR market overperformance relative to the CDA market. All data are from Season 3.

this evidence is correlational and the correlation itself is not strong, these results are directionally consistent with Hanson's expectations that LMSR markets would outperform CDA markets when there are few active traders.

Another possible source of CDA market underperformance relates to the timing within a question. Traders may be reluctant to place many large orders when question resolutions and expected payoffs are months into the future, and instead may prefer to allocate their attention and artificial currency to questions with more imminent resolutions. In CDA markets specifically, low activity on a question at a given time corresponds to thinner order books, which in turn makes these questions less attractive to other traders. LMSR markets, on the other hand, should be less vulnerable to such negative activity feedback loops since traders can always trade with the automated market maker.

We performed two tests on whether the differences in accuracy vary with time within a question. The first uses the absolute time until the question is resolved, the second uses the proportion of time until the question is resolved. The left side of Figure 4 shows Brier scores by absolute timing within question. Time is defined as the number of days to question resolution. For example, we may compare the difference in accuracy 200 days versus 100 days versus 1 day before question resolution. The lines represent linear model fits and include data points across all available questions at the given point in time (e.g., all questions open 100 days before resolution). As shown in the figure, for this specification, the LMSR market tended to produce more accurate forecasts throughout the duration of the questions and differences in accuracy were relatively constant. A mixed-effects model with

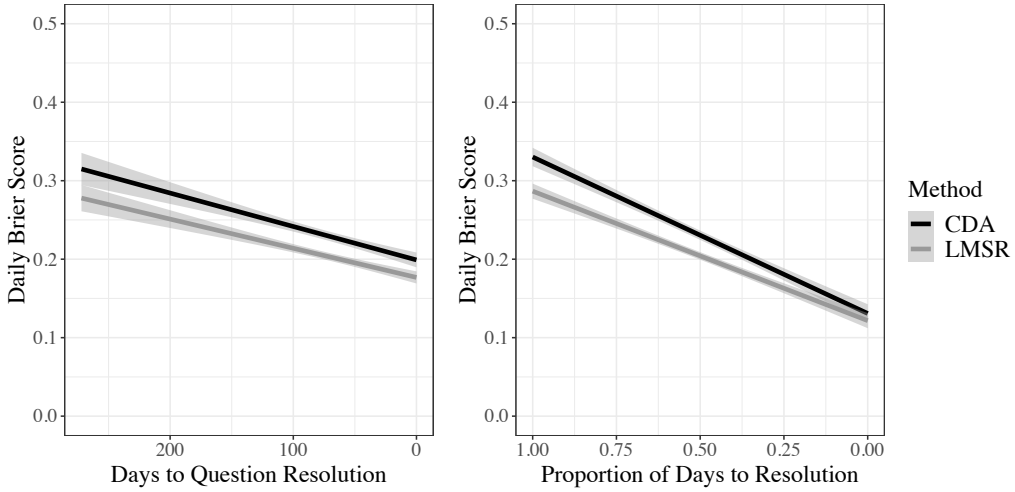

Figure 4: Left side: Brier scores for CDA and LMSR markets by the absolute number of days to resolution within a question. For example, 100 on the horizontal axis denotes that that the scores are measured 100 days before question resolution. Right side: Brier scores for CDA and LMSR markets by the proportion of days to resolution within a question, where 1 denotes question start and 0 denotes time of question resolution. Linear model fits shown for each market on both left and right side.

question-level random intercepts is consistent with the visual result revealing no significant interaction between prediction market type and absolute time ($b = -.005, p = .11$). (See Table 10, Column B.)

On the right side of Figure 4, we normalize timing by question duration, using the proportion of days to resolution instead of absolute number of days. For example, for questions lasting 100 days, we may compare the difference in accuracy at the start of the questions (1.00), at day 50, which is in the middle of question duration (0.50), and right before question resolution. When using normalized timing, we see convergence over time. Whereas the CDA market tends to underperform in the early periods within a question, the two markets

are almost tied in accuracy immediately before questions resolve. A linear mixed-effects model yields results consistent with convergence, revealing a significant interaction between market type and relative question timing ($b = -.034, p < .001$), denoting that the difference in Brier scores moved in favor of CDA markets by 0.034 Brier score points between the start and the end of a question's duration. (See Table 10, Column C.)

Overall, the exploratory analysis of within-question timing suggests that the CDA market underperforms the LMSR market by a margin that is most pronounced in the early stages of questions.

3.3. Complementary Research Question 2: Identifying Accurate Forecasters

The most notable result from our main investigation is the large impact of forecasters with superior accuracy track records relative to the choice of crowd prediction system. This underscores the value of reliable measures and methods to identify those highly accurate individuals. Are prediction polls better at reliably identifying accurate forecasters than prediction markets? We answer the question separately for sub-elite and elite forecasters.

In our sub-elite reliability analysis, we calculate percentile ranks within condition for each season. The sample includes all sub-elite forecasters who were active in two subsequent seasons. Rankings are based on Brier scores for prediction polls and season-end earnings for prediction markets. Using these percentile ranks, we calculate the correlation of within-condition ranks across seasons. A test-retest Pearson product-moment correlation of $r = 1$ means that the percentile rank of a forecaster in one season would perfectly predict their rank in the next season, while r = 0 means that rankings between seasons are completely independent of one another.

Table 11: Test-retest reliability of sub-elite forecaster rankings in prediction polls and prediction markets.

Period	Prediction Markets		Prediction Polls		Difference, p-value
	n	r	n	r	p
Seasons 2–3	237	0.25	412	0.37	.101
Seasons 3–4	508	0.18	244	0.44	$< .001$
Seasons All	591	0.20	549	0.38	$< .001$

Table 12: Season 3 performance among N = 122 elite forecasters as a function of forecaster condition in Season 2, for three accuracy measures: A. Raw Brier Score, B. Final Rank, and C. Standardized Brier Score. OLS regression models coefficients reported, with standard errors in parentheses.

Season 2 Condition	A. Raw BS	B. Rank	C. Std. BS
Intercept	0.196 (0.006)	0.57 (0.06)	0.23 (0.07)
Indep. Polls (Ref.)			
Team Polls	-0.003 (0.009)	0.00 (0.08)	-0.10 (0.10)
Markets	-0.005 (0.009)	-0.01 (0.08)	-0.17 (0.10)
Superpolls	-0.025 (0.008) **	-0.17 (0.07) *	-0.34 (0.08) **
Adj. R-squared	0.10	0.05	0.13

Note: * $p < .05$, ** $p < .01$. Lower values denote better performance for forecasters from a given condition relative to the reference group of independent poll forecasters.

Table 11 shows the results separated by prediction system (prediction markets versus prediction polls) and season pairs (Seasons 2–3 and Seasons 3–4). All correlation coefficients are significantly higher than zero ($t > 4.00, p < .001$). For Seasons 2–3, prediction polls produce more reliable rankings ($r = 0.37$, $n = 412$, $p < .001$) than prediction markets ($r = 0.25$, $n = 237$, $p < .001$). The difference between the two correlation coefficients is not significant ($z = 1.64$, $p = .101$). The difference in test-retest reliability is more pronounced in Seasons 3–4 (polls: $r = 0.44$, $n = 244$, $p < .001$; markets: $r = 0.18$, $n = 508$, $p < .001$; difference: $z = 3.71$, $p < .001$). In an overall analysis, we combine data across seasons, using the first season pair available for each forecaster. For example, if a forecaster is active in Seasons 2, 3 and 4, we would only include their data from Seasons 2 and 3. We find prediction polls produced more reliable performance rankings than prediction markets.

To assess each system's reliability in identifying elite forecasters, we used data from Season 3, in which all elite forecasters competed in (team-based) Superpolls. Forecasters had attained elite status by placing in the top 2% of their randomly assigned condition (prediction markets or prediction polls) in Seasons 1 or 2. We compared the season-end individual accuracy of these newly identified elite forecasters in prediction polls versus prediction markets, using raw Brier scores as the primary performance measure. See Table 12, Column A. Season-end leaderboard rank and z-score-standardized Brier scores were used in sensitivity analyses. There were no significant differences in Season 3 raw Brier scores among newly-qualified elite forecasters based on their prior condition. Elite forecasters who qualified from team-

based polls and prediction markets earned similar scores, relative to each other and to those from independent polls. The only group that achieved significantly better Brier scores than the reference group were returning elite forecasters from Season 2 ($b = -0.025$, $se = 0.008$, $p < .01$). Sensitivity analyses with the alternative performance measures yielded similar results. See Table 12, Columns B and C.

Overall, there was no evidence that top prediction market traders underperformed forecasters who had attained elite status in either independent or team prediction polls, when all competed in Superpolls. These results suggest that prediction markets are approximately as effective in identifying highly accurate forecasters as prediction polls. This finding is inconsistent with our expectations with respect to Complementary Research Question 2.

4. Discussion

4.1. Research Implications

Our key result for our main research question is that the superior aggregate accuracy of elite forecasters holds regardless of whether the forecasters are working in prediction markets or prediction polls, and whether more or less sophisticated aggregation algorithms are used in polls. Analyses for our first complementary research question show that, in our setting, featuring several hundred traders working on 100+ forecasting questions, LMSR markets yield more accurate aggregate predictions than CDA markets. Our second complementary research question examines skill identification and reveals that elite forecasters perform similarly, independent of whether they were originally identified in prediction polls or prediction markets.

The strong performance of elite crowds in both prediction markets and prediction polls underscores the value of designing and improving methods to identify such high-performers early and reliably. Technology choices relevant to talent spotting include the choice of performance tracking methods (Gneiting and Raftery 2007, Witkowski et al. 2017), behavioral data capture (Atanasov et al. 2020, Karvetski et al. 2022, Atanasov and Himmelstein 2023), and aligning forecaster incentives (Lichtendahl et al. 2013, Witkowski et al. 2023). More generally, our results highlight one important aspect of system design: the choice of which individuals are granted access may have a larger impact on the system's overall performance than any other feature.

Our main result regarding the comparison of systems, a virtual tie between LMSR prediction markets and team-based prediction polls in terms of accuracy, differs from that of Atanasov et al. (2017), who reported that team-based prediction polls significantly outperformed CDA prediction markets. The results of our complementary market comparison explain this seeming discrepancy across studies since we show that LMSR prediction markets tend to produce more accurate forecasts than CDA prediction markets. To summarize GJP results across studies and seasons, team-based prediction polls and LMSR prediction markets tend to yield similar levels of accuracy, while CDA markets produce somewhat less accurate aggregate forecasts. The difference in accuracy between LMSR and CDA markets is largely traceable to questions attracting few traders and to early periods within a question. This result is directionally consistent with Hanson's theory-based prediction that, relative to CDA markets, LMSR markets will perform especially well

in thin-market conditions.

The present work extends the literature on forecaster tracking from prediction polls to prediction markets. We find that prediction markets with elite traders tend to outperform those populated by less selective crowds. LMSR prediction markets populated exclusively with elite traders (Supermarkets) produce similar levels of accuracy as Superpolls, team-based prediction polls populated with only elite forecasters. The latter remain unbeaten in formal comparisons.

4.2. Practical Implications

The practical question we set to address focused on a manager who seeks to maximize forecasting performance in a crowdsourcing environment through her choices about forecasting systems and crowds. Our investigation points to specific recommendations.

The first choice concerns system type. The three systems we examine are prediction polls, CDA prediction markets, and LMSR prediction markets. Our study featured crowds of hundreds of forecasters and 100 to 150 questions over roughly 9 months. With these parameters, we find that CDA markets underperform LMSR markets, which are in turn tied in accuracy with prediction polls. Thus, when the ratio of sub-elite forecasters to questions is roughly 5 to 1 or lower, the manager should generally avoid the CDA market structure.

Regarding the choice between LMSR prediction markets and prediction polls, our results point to an approximate tie in terms of accuracy. If the manager has run a prediction market for years, and is generally satisfied with its usability, they may avoid system switching costs. If, however, the crowd-

sourcing initiative is new and managers are interested not just in aggregate accuracy, but in identifying reliably accurate forecasters, they may be better served by employing a prediction poll.

Second, our results offer a clear recommendation for improving accuracy: employ smaller, elite crowds. These findings are relevant to corporate forecasting tournaments (Cowgill and Zitzewitz 2015) as well as to the growing research literature on public forecasting tournaments (Tetlock et al. 2017, Morstatter et al. 2019, Atanasov et al. 2023). Whether the prediction system is an LMSR market or prediction polls, managers could improve performance by selecting a smaller, elite crowd based on prior performance in the competition.

Small, elite forecaster crowds may yield benefits beyond accuracy. For example, when forecasts use proprietary data or relate to confidential outcomes, employing a smaller group of forecasters may help minimize information leakage. This is a non-trivial concern, especially in prediction markets: Google Chief Economist Hal Varian has noted that data concerns were key to stopping one Google prediction markets project: "The problem is, the things that we really wanted to get a probability assessment on were things that were so sensitive that we thought we would violate the SEC rules on insider knowledge because [...] anybody who looks at the auction is now an insider" (Cowen and Varian 2019). [4]

Finally, the formation of elite forecasting pools depends on picking rea-

[4]Improvements in confidentiality can be achieved either by employing smaller, elite crowds or by deploying prediction polls instead of prediction markets. Polls can function well without broadcasting crowd consensus to all active forecasters (Atanasov et al. 2017).

sonable performance cutoffs. In our data, the top 2% of forecasters were deemed elite. This cutoff decision was partly driven by the details of the tournament, such as the number of forecasting questions (100–150 per season) and number of forecasters (1,000 to 3,000). Thus, it should not be considered a hard-and-fast rule. A rough order-of-magnitude recommendation based on the previous literature is that at least 5-10 forecasters should be available to answer each question (Mannes et al. 2014). The results of our sub-sampling simulations in Superpolls suggest that active, elite crowds produce accurate aggregate forecasts even when they are very small in number, with forecaster-to-question ratios of 1:5, e.g., 20 forecasters for 100 questions over a 9-month period. We note that the average elite forecaster in our sample answered more than 70% of available questions and made over 5 forecasts per question, so these results on very small crowds depend on the availability of highly engaged forecasters. This implies that raising the threshold for entry to elite status to the top 1% may work well when sourcing from a crowd of 2,000+ forecasters. On the other hand, our results on the high reliability in performance across seasons, especially in prediction polls, implies that moderately relaxing the threshold for promotion to elite status (e.g., from top 2% to top 5%) would result in including additional high performers, and is thus unlikely to materially reduce aggregate performance.

4.3. Limitations and Future Directions

As is true for virtually all empirical investigations of individual and system-level performance, the current results should be generalized with caution. First, this study featured probabilistic forecasting questions on geopolitics, economics, and public health. Questions were designed to be rigorously

resolvable in the near future. Future work should also explore the merits of small, elite crowds in prediction markets and prediction polls for answering other types of questions, such as ones about long-term trends or rare events. More generally, while the two widely used crowd prediction systems in our comparison produced similar levels of performance, it is possible that novel systems will perform substantially better or worse than our comparison set. Thus, our results should not be seen as a general statement that the choice of prediction system does not matter for maximizing forecasting performance.

Finally, the comparison between Supermarkets and Superpolls did not feature random assignment. Most notably, returning elite forecasters self-selected into a system of their choice, and most chose to remain in the Superpoll system they had worked in. Relative to random assignment, this preference may have provided a small benefit to Superpolls, though it is unlikely to have qualitatively changed the result: a statistical tie between Supermarkets and Superpolls. Superpolls exhibited a small and insignificant advantage over Supermarkets, and a large swing, equivalent to a 25% change in relative Brier scores, would have been needed to produce a significant advantage of Supermarkets over Superpolls.

4.4. Conclusion

This study is the first to demonstrate that small, elite crowds outperform large, less selective crowds across two popular prediction systems: prediction markets and prediction polls. This finding underscores the more general point that the performance of information systems utilizing human inputs depends crucially on the humans making those inputs. In the context of prediction systems, the main challenge remains the identification of elite forecasters

with little or no historical performance data. We believe the development of methods addressing this challenge will be a fruitful area for future research.

Acknowledgement

This research was supported by the Open Philanthropy Foundation and by a research contract from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center, contract number 140D0419C0049. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Atanasov, P., Himmelstein, M., 2023. Talent spotting in crowd prediction, in: Seifert, M (ed.), Judgment in Predictive Analytics. Springer, pp. 135–184.

Atanasov, P., Rescober, P., Stone, E., Swift, S.A., Servan-Schreiber, E., Tetlock, P., Ungar, L., Mellers, B., 2017. Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. Management Science 63, 691–706.

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., Tetlock, P., 2020. Small steps to accuracy: Incremental belief updaters are better forecasters. Organizational Behavior and Human Decision Processes 160, 19–35.

Atanasov, P.D., Joseph, R., Feijoo, F., Marshall, M., Siddiqui, S., 2023. Human Forest vs. Random Forest in Time-Sensitive COVID-19 Clinical Trial Prediction. Working Paper.

Berg, J.E., Rietz, T.A., 2003. Prediction markets as decision support systems. Information systems frontiers 5, 79–93.

Brier, G.W., 1950. Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review 78, 1–3.

Budescu, D.V., Chen, E., 2015. Identifying Expertise to Extract the Wisdom of Crowds. Management Science 61, 267–280.

Cowen, T., Varian, H., 2019. Hal Varian on Taking the Academic Approach to Business (Ep. 69). https://conversationswithtyler.com/episodes/hal-varian/. [Online; accessed 18-December-2021].

Cowgill, B., Zitzewitz, E., 2015. Corporate Prediction Markets: Evidence from Google, Ford, and Firm X. The Review of Economic Studies 82, 1309–1341.

Davis-Stober, C.P., Budescu, D.V., Dana, J., Broomell, S.B., 2014. When Is a Crowd Wise? Decision 1, 79–101.

Forsythe, R., Nelson, F., Neumann, G.R., Wright, J., 1992. Anatomy of an Experimental Political Stock Market. The American Economic Review , 1142–1161.

Gneiting, T., Raftery, A.E., 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association 102, 359–378.

Goldstein, D.G., McAfee, R.P., Suri, S., 2014. The Wisdom of Smaller, Smarter Crowds, in: Proceedings of the 15th ACM Conference on Economics and Computation (EC'14), pp. 471–488.

Goldstein, S., Hartman, R., Comstock, E., Baumgarten, T.S., 2016. Assessing the accuracy of geopolitical forecasts from the US Intelligence Community's prediction market. Working Paper.

Good Judgment Project, 2016. GJP Data. URL: https://doi.org/10.7910/DVN/BPCDH5, doi:10.7910/DVN/BPCDH5.

Graefe, A., Armstrong, J.S., 2011. Comparing face-to-face meetings, nominal groups, delphi and prediction markets on an estimation task. International journal of forecasting 27, 183–195.

Hanson, R., 2003. Combinatorial Information Market Design. Information Systems Frontiers 5, 107–119.

Hanson, R., 2007. Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation. The Journal of Prediction Markets 1, 3–15.

Healy, P.J., Linardi, S., Lowery, J.R., Ledyard, J.O., 2010. Prediction Markets: Alternative Mechanisms for Complex Environments with Few Traders. Management Science 56, 1977–1996.

Jose, V.R.R., Nau, R.F., Winkler, R.L., 2009. Sensitivity to Distance and Baseline Distributions in Forecast Evaluation. Management Science 55, 582–590.

Kahneman, D., Sibony, O., Sunstein, C.R., 2021. Noise: A Flaw in Human Judgment. William Collins Publishers.

Karvetski, C.W., Meinel, C., Maxwell, D.T., Lu, Y., Mellers, B.A., Tetlock, P.E., 2022. What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters? International Journal of Forecasting 38, 688–704.

Lichtendahl, K.C., Grushka-Cockayne, Y., Pfeifer, P.E., 2013. The Wisdom of Competitive Crowds. Operations Research 61, 1383–1398.

Malkiel, B.G., Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. The Journal of Finance 25, 383–417.

Mannes, A.E., Soll, J.B., Larrick, R.P., 2014. The Wisdom of Select Crowds. Journal of Personality and Social Psychology 107, 276–299.

Markose, S.M., 2005. Computability and evolutionary complexity: markets as complex adaptive systems (cas). The Economic Journal 115, F159–F192.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S.E., Ungar, L., Bishop,

M.M., Horowitz, M., Merkle, E., Tetlock, P., 2015a. The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics. Journal of Experimental Psychology: Applied 21, 1–14.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., et al., 2015b. Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. Perspectives on Psychological Science 10, 267–281.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S.E., Moore, D., Atanasov, P., Swift, S.A., et al., 2014. Psychological Strategies for Winning a Geopolitical Forecasting Tournament. Psychological Science 25, 1106–1115.

Morstatter, F., Galstyan, A., Satyukov, G., Benjamin, D., Abeliuk, A., Mirtaheri, M., Hossain, K.T., Szekely, P.A., Ferrara, E., Matsui, A., et al., 2019. Sage: A hybrid geopolitical event forecasting system., in: 28th International Joint Conference on Artificial Intelligence (IJCAI'19), pp. 6557–6559.

Murphy, A.H., Winkler, R.L., 1987. A General Framework for Forecast Verification. Monthly Weather Review 115, 1330–1338.

Page, L., Clemen, R.T., 2013. Do Prediction Markets Produce Well-Calibrated Probability Forecasts? The Economic Journal 123, 491–513.

Page, S.E., 2007. Making the Difference: Applying a Logic of Diversity. Academy of Management Perspectives 21, 6–20.

Peeters, T., 2018. Testing the Wisdom of Crowds in the field: Transfermarkt valuations and international soccer results. International Journal of Forecasting 34, 17–29.

Pennock, D., 2006. Implementing Hanson's Market Maker.

http://blog.oddhead.com/2006/10/30/implementing-hansons-market-maker/. Accessed: 2021-12-18.

Polgreen, P.M., Nelson, F.D., Neumann, G.R., Weinstein, R.A., 2007. Use of Prediction Markets to Forecast Infectious Disease Activity. Clinical Infectious Diseases 44, 272–279.

Reade, J.J., Williams, L.V., 2019. Polls to probabilities: Comparing prediction markets and opinion polls. International Journal of Forecasting 35, 336–350.

Rothschild, D.M., Sethi, R., 2016. Trading Strategies and Market Microstructure: Evidence from a Prediction Market. The Journal of Prediction Markets 10, 1–29.

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

Schmitz, J., Rothschild, D., 2019. Understanding market functionality and trading success. Plos One 14, e0219606.

Servan-Schreiber, E., Wolfers, J., Pennock, D.M., Galebach, B., 2004. Prediction Markets: Does Money Matter? Electronic Markets 14, 243–251.

Silver, N., 2012. The Signal and the Noise: Why So Many Predictions Fail-but Some Don't. Penguin Publishing Group.

Snowberg, E., Wolfers, J., Zitzewitz, E., 2013. Prediction Markets for Economic Forecasting, in: Handbook of Economic Forecasting. Elsevier. volume 2. chapter 11, pp. 657–687.

Spann, M., Skiera, B., 2003. Internet-Based Virtual Stock Markets for Business Forecasting. Management Science 49, 1310–1326.

Strijbis, O., Arnesen, S., 2019. Explaining variance in the accuracy of prediction markets. International Journal of Forecasting 35, 408–419.

Surowiecki, J., 2005. The Wisdom of Crowds. Anchor.

Tetlock, P.E., 2005. Expert Political Judgment: How Good Is It? How Can We Know? Princeton University Press.

Tetlock, P.E., Gardner, D., 2016. Superforecasting: The Art and Science of Prediction. Random House.

Tetlock, P.E., Mellers, B.A., Scoblic, J.P., 2017. Bringing probability judgments into policy debates via forecasting tournaments. Science 355, 481–483.

Winkler, R.L., 1968. The Consensus of Subjective Probability Distributions. Management Science 15, B–61–B–75.

Witkowski, J., Atanasov, P., Ungar, L.H., Krause, A., 2017. Proper Proxy Scoring Rules, in: Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17), pp. 743–749.

Witkowski, J., Freeman, R., Wortman Vaughan, J., Pennock, D.M., Krause, A., 2023. Incentive-Compatible Forecasting Competitions. Management Science , Forthcoming.

Wolfers, J., Zitzewitz, E., 2004. Prediction Markets. Journal of Economic Perspectives 18, 107–126.

Woolley, A.W., Chabris, C.F., Pentland, A., Hashmi, N., Malone, T.W., 2010. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. Science 330, 686–688.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.