Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls

Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, Barbara Mellers

To cite this article:
Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, Barbara Mellers (2017) Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. Management Science 63(3):691-706. http://dx.doi.org/10.1287/mnsc.2015.2374

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 3, March 2017, pp. 691–706
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls

Pavel Atanasov,[a, b] **Phillip Rescober,**[c] **Eric Stone,**[d] **Samuel A. Swift,**[e] **Emile Servan-Schreiber,**[f] **Philip Tetlock,**[g] **Lyle Ungar,**[h] **Barbara Mellers**[i]

[a] University of Pennsylvania, Philadelphia, Pennsylvania 19104; [b] Polly Portfolio, Inc., New York, New York 10007; [c] Good Judgment, Inc., Philadelphia, Pennsylvania 19103; [d] Pluralsight, Inc., Boston, Massachusetts 02118; [e] Betterment, Inc., New York, New York 10010; [f] Lumenogic, Inc., 75013 Paris, France; [g] Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [h] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [i] Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Contact: pdatanasov@gmail.com (PA); rescober@goodjudgment.com (PR); theericstone@gmail.com (ES); samswift@gmail.com (SAS); emile@lumenogic.com (ES-S); tetlock@wharton.upenn.edu (PT); ungar@cis.upenn.edu (LU); mellers@wharton.upenn.edu (BM)

Received: September 14, 2014
Revised: March 11, 2015; June 24, 2015
Accepted: September 14, 2015
Published Online in Articles in Advance: April 22, 2016

https://doi.org/10.1287/mnsc.2015.2374

Abstract. We report the results of the first large-scale, long-term, experimental test between two crowdsourcing methods: prediction markets and prediction polls. More than 2,400 participants made forecasts on 261 events over two seasons of a geopolitical prediction tournament. Forecasters were randomly assigned to either prediction markets (continuous double auction markets) in which they were ranked based on earnings, or prediction polls in which they submitted probability judgments, independently or in teams, and were ranked based on Brier scores. In both seasons of the tournament, prices from the prediction market were more accurate than the simple mean of forecasts from prediction polls. However, team prediction polls outperformed prediction markets when forecasts were statistically aggregated using temporal decay, differential weighting based on past performance, and recalibration. The biggest advantage of prediction polls was at the beginning of long-duration questions. Results suggest that prediction polls with proper scoring feedback, collaboration features, and statistical aggregation are an attractive alternative to prediction markets for distilling the wisdom of crowds.

History: Accepted by Uri Gneezy, behavioral economics.
Funding: This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center (DoI/NBC) [Contract D11PC20061].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2015.2374.

Keywords: forecasting • prediction markets • crowdsourcing, belief elicitation

1. Introduction

We examine two crowd-based approaches for collecting predictions about future events: prediction markets and prediction polls. In an idealized prediction market, traders are motivated by market profits to buy and sell shares of contracts about future events. If and when they obtain relevant information, they act quickly in the market. Knowledge is continuously updated and aggregated, making prices generally good estimates of the chances of future events. In the strong form of the efficient markets hypothesis, no additional information should be able to improve the accuracy of the last price. This form of crowdsourcing is used in many organizations, bridging the gap between economic theory and business practices (Cowgill and Zitzewitz 2015, Spann and Skiera 2003, Surowiecki 2005).

Direct probability elicitation methods are quite different from prediction markets and are often misunderstood. We examine a version of probability elicitation that we call prediction polls. Participants offer probabilistic forecasts, either independently or as members of a team, and update their beliefs throughout the duration of each question as often as they wish. They receive feedback about their performance using a proper scoring rule and have many opportunities for learning. Prediction polls should not be confused with opinion polls. In opinion polls, respondents are typically asked about their personal preferences or intentions, on a single occasion. In prediction polls, respondents place predictions on future events, aiming for accuracy. Past research shows that that participants' predictions of election outcomes provide more accurate estimates than participants' stated preferences about election outcomes (Rothschild and Wolfers 2010).

What is the value of better predictions? The answer varies across businesses, governments, and individuals, but revenues in the forecasting industry are estimated at $300 billion in current dollars (Sherden 1998). Better predictions are financially consequential. Thus, it is important to know which methods provide more accurate estimates. Prediction markets are often used by organizations to answer questions about geopolitical events, project completion, or product sales. But, in this context, prediction polls are less common. There is little experimental evidence on the relative accuracy of alternative methods. The current paper addresses this knowledge gap and provides lessons about how to improve the accuracy of crowdsourced predictions more generally.

There are theoretical and empirical reasons to believe that prediction markets and prediction polls are both valid means of gathering and aggregating information (Gürkaynak and Wolfers 2006, Hanson 2003). These methods share essential properties: they account for information revelation over time, place higher weights on better forecasters, and use past forecasts to improve calibration. We put prediction markets and prediction polls in a head-to-head competition and measured the accuracy of forecasts they produced over two years of a geopolitical forecasting tournament.

2. Background and Hypotheses

An opportunity to test the accuracy of prediction markets against prediction polls arose in 2011 when the Intelligence Advanced Research Project Agency (IARPA), the research wing of the intelligence community, sponsored a multiyear forecasting tournament. Five university-based programs competed to develop the most innovative and accurate methods possible to predict a wide range of geopolitical events. Will any country officially announce its intention to withdraw from the Eurozone before April 1, 2013? Will the six-party talks on the Korean Peninsula resume before January 1, 2014? What will be the lowest end-of-day price of Brent Crude Oil between October 16, 2013, and February 1, 2014? Our group experimentally tested a variety of elicitation and aggregation methods. Our methods produced highly accurate forecasts and helped our team, the Good Judgment Project, to win the forecasting tournament (see the online appendix).

The current paper offers the most thorough analysis of prediction market data from the Good Judgment Project, as well as a detailed description of aggregation challenges and solutions in prediction polls. Our main focus is the comparison of prediction markets and prediction polls.

We build on our understanding of individual forecasting behavior described in prior work. Mellers et al. (2014) provide an overview of the forecasting tournament and discuss the positive impacts of three behavioral interventions—training, teaming, and tracking—on individual performance in prediction polls. Mellers et al. (2015a) explore the profiles of individuals using dispositional, situational, and behavioral variables. Mellers et al. (2015b) document the performance of the most accurate performers, known as superforecasters, and present four complementary reasons for their success. Aggregation techniques for prediction poll data are discussed in three papers. Satopää et al. (2014a) offer a simple method for combining probability estimates in log-odds space. The method uses differential weighting, discounting, and recalibration or “extremizing” of forecasts to reflect the amount of overlapping information of individual opinions. Satopää et al. (2014b) describe a time-series model for combining expert estimates that are updated infrequently. Baron et al. (2014) provide a theoretical justification and empirical evidence in favor of transforming aggregated probability predictions toward the extremes. Atanasov et al. (2013) develop a method for aggregating probability estimates in prediction markets when probabilities are inferred from individual market orders and combined using statistical aggregation approaches. Finally, Tetlock et al. (2014) discuss the role that tournaments can take in society by both increasing transparency and improving the quality of scientific and political debates by opening closed minds and increasing assertion-to-evidence ratio.

We begin by describing prediction markets and prediction polls. We then discuss the aggregation challenges inherent in dynamically elicited probability forecasts. Next, we offer hypotheses and test them in two studies covering different seasons of the tournament. We also examine the sensitivity of our results to choice of aggregation parameters, scoring rules, and potential moderators.

2.1. Prediction Markets

Since the work of Hayek (1945), economists have championed the power of markets as a mechanism for aggregating information that is widely dispersed among economic actors. The efficient markets hypothesis posits that the latest price reflects all information available to market participants, so future price movements are unpredictable without insider knowledge (Fama 1970).

Prediction markets build on this rich intellectual tradition and are designed to produce continuously updated forecasts about uncertain events. In binary options markets, traders place bets on the probability of event occurrence. For example, an election contract may pay $1 if candidate A is elected president, and $0 otherwise. The latest price is presumed to reflect the current best guess about the probability of the event. A contract trading at 60 cents on the dollar implies that market participants, in aggregate, believe the event is 60% likely to occur. Those who think the current price is a poor estimate of the event probability have incentives to trade.

Evidence suggests that prediction markets can outperform internal sales projections (Plott and Chen 2002), journalists’ forecasts of Oscar winners (Pennock et al.

2001), and expert economic forecasters (Gürkaynak and Wolfers 2006). Furthermore, prediction markets perform at least as well as aggregated opinion polls of voter preferences (Berg et al. 2001, Rothschild 2009).

Some researchers have noted systematic distortions in market prices, including the favorite long-shot bias (Page and Clemen 2012) and partition dependence (Sonnemann et al. 2013). Theoretical work has described conditions under which prediction markets fail to incorporate individual beliefs (Manski 2006, Ostrovsky 2012). However, these objections have not changed the overall impression that prediction markets are superior to most other prediction methods. Such results have led Wolfers (2009) to recommend that "rather than averaging forecasts of ... [economic forecasters, they] should be put in a room and told to bet or trade against each other" (p. 38).

In the current study, several hundred participants made trades in a continuous double auction (CDA) market. In a CDA, buyers and sellers are matched to determine the market price. Traders place orders in the form of bids (buy orders) and asks (sell orders), and records are kept in an order book. Trades are executed when an order is placed and the highest buying price matches or exceeds the lowest selling price. Rules of the tournament prohibited us from using real monetary incentives so all transactions were made using hypothetical currency.

2.2. Prediction Polls

In prediction polls, forecasters express their beliefs by answering the question, "How likely is this event?" Probability judgments are validated against ground truth using a proper scoring rule. Forecasters receive feedback on their performance, in much the same way that prediction market traders are ranked based on earnings. A similar method has appeared in the literature under the name of *competitive forecasting* (Servan-Schreiber 2007).

There are two important distinctions between prediction polls and other polls or surveys. First, in prediction polls, participants are asked for probabilistic forecasts, rather than preferences or voting intentions. Forecasts are elicited in a dynamic context. Forecasters update their predictions whenever they wish, and feedback is provided when events are resolved. Second, forecasters compete against other forecasters. Servan-Schreiber (2012) has argued that competitive features encourage better search processes and more accurate inferences. Third, prediction polls rely on crowds with dozens, hundreds, or thousands of individuals who may be knowledgeable but are not necessarily subject matter experts, which distinguishes polls from expert elicitation techniques.

Our team examined two forms of prediction polls. In the independent poll condition, forecasters worked independently and had no access to the opinions of others. In the team poll condition, forecasters worked online in teams of approximately 15 members (see Mellers et al. 2014). They shared information, discussed rationales, and encouraged each other to forecast. Teams did not need to reach consensus; they made individual predictions, and the median was defined as the team forecast. Individual leader boards (within teams and across independent forecasters) displayed Brier scores. Team leader boards (across teams) displayed group performance using Brier scores. Forecasters in team prediction polls reliably outperformed forecasters in independent prediction polls.

2.3. Comparisons of Markets and Polls

Studies comparing prediction markets and methods similar to prediction polls have used observational field data or small-scale laboratory experiments. Most have focused on short-term sports predictions or general knowledge tests. In contrast, we examine geopolitical predictions on questions that are open for several months, on average.

Chen et al. (2005) compared the accuracy of linear and logarithmic aggregations of polls[1] to prices in a test based on final scores of football games. Accuracy of the methods did not differ using the absolute distance rule, the quadratic scoring rule, or the logarithmic scoring rule.[2] Goel et al. (2010) compared real-money prediction markets to probabilities from opinion polls (ProbabilitySports and Amazon's Mechanical Turk). Prediction markets were more accurate, but not by a significant margin.

Rieg and Schoder (2010) found no differences in accuracy when comparing markets and opinion polls in small-scale experiments, where each of 6 to 17 participants made a single prediction in either a market or a poll. In another small-group experiment, Graefe and Armstrong (2011) compared the accuracy of answers to general knowledge questions using the Delphi method, the Nominal Groups technique, and a prediction market with a logarithmic market scoring rule (LMSR) (Hanson 2003). The accuracy of the methods was approximately equal. In summary, prior research has found polls and markets to be closely matched in accuracy. None of the existing studies simultaneously featured experimental assignment and evaluation on a large set of long-term, real-world questions.

2.4. Forecast Aggregation in Prediction Polls

An important challenge for dynamic forecasting is aggregating forecasts over time (Armstrong 2001). We discuss a parsimonious and effective aggregation algorithm, which our team devised to address three challenges—outdated predictions, heterogeneity of individual skills, and miscalibration—discussed below.

(1) *Outdated Predictions*. When forecasts are continuously elicited and updated over time, more recent ones usually contain fresher, more valuable information. For example, one-day-ahead temperature forecasts are highly accurate, whereas those issued 10 days prior are as accurate as predictions based purely on historical averages (Silver 2012).

However, although the single most recent forecast is likely to be closer to the truth than an older one, just one forecast can be noisy and inaccurate. The challenge is to strike the right balance between recency and number of forecasts. In prediction markets, the last price is treated as the crowd's best estimate. In CDA markets, two marginal traders produce the price: the most recent highest bidder and lowest-price seller, although, in highly liquid markets these two individuals have limited ability to influence the price. If traders had chosen to transact at different prices, others could have jumped in to set the price. The availability of price history and the depth of the order book make it easier to maintain stable prices.

In prediction polls, forecasters have incentives to update their predictions, but they do not all update at the same time. A simple mean relies too much on older predictions, making it unresponsive to new developments. One solution is exponential discounting, where the weight of each estimate depends on the length of time since it was last updated. This approach is a moving average of a predefined number (or proportion) of the most recent forecasts. For example, the algorithm might use the latest 30 updates or the most recent 20% of updates.

(2) *Heterogeneity of Individual Skills*. Prediction markets automatically provide a mechanism for incorporating differences in forecaster knowledge and skill. In the short run, the effect of an order depends on the number of shares that a trader wants to buy or sell. Order size can be construed as a measure of one's confidence in a position. In finite markets,[3] larger orders are more influential. For a rational, risk-averse trader, the amount invested in a given contract should be proportional to the difference between the trader's private beliefs and the market price (Wolfers and Zitzewitz 2006). In the long run, traders who place correct bets tend to be rewarded with higher earnings, and their increased wealth affords them greater power to influence future prices. This feature of markets is central to the marginal trader hypothesis, which stipulates that the efficiency of prediction markets is driven by a minority of unbiased and active participants who wield corrective influence (Forsythe et al. 1992).

If some forecasters in prediction polls are consistently better than others, aggregation algorithms can improve performance by placing greater weight on the predictions of forecasters who are more likely to be accurate. We found that prior accuracy and belief updating history were robustly indicative of future performance and incorporated these into our aggregation algorithms.

(3) *Miscalibration*. Markets provide incentives for individuals to correct systematic biases, such as overconfidence or underconfidence. A rational trader would place bets that are profitable in expectation, realigning prices with historical base rates. Prediction markets generally produce adequately calibrated prices, with the exception of the favorite long-shot bias.

Despite the well-known tendency for individuals to be overconfident when making probability judgments, simple aggregates of prediction polls tend to be underconfident (Baron et al. 2014, Satopää et al. 2014b). In other words, polls tend to produce average forecasts that are too close to the ignorance prior, e.g., 50% for a two-option question. Baron et al. (2014) discuss two reasons for aggregate forecasts appearing to be underconfident. First, because the probability scale is bounded, noise in individual estimates tends to push mean probability estimates away from the extremes. This challenge is partially addressed by using the median rather than the mean to aggregate predictions. Second, although each individual may possess only a portion of knowable information, the aggregate is generally based on more information than any single forecast. The tendency of average predictions to be better informed but less extreme than individual estimates makes recalibration beneficial.

2.5. Hypotheses

We offer two hypotheses about the relative accuracy of methods. The first is driven by the importance of outdated forecasts, inattention to past performance, and underconfident estimates in prediction polls.

Hypothesis 1. *Prediction markets will outperform prediction polls when forecasts are combined using a simple mean.*

The second hypothesis reflects the ability of statistical algorithms to address these aggregation challenges. Prior research also suggests that teaming confers a consistent advantage to the accuracy of polls Mellers et al. (2014).

Hypothesis 2. *Prediction polls will outperform prediction markets when probability estimates are elicited in teams and combined with algorithms featuring decay, differential weighting, and recalibration.*

3. Experiment 1: Methods

The Aggregative Contingent Estimation (ACE) tournament was sponsored by IARPA and lasted for four years (from 2011 to 2015). The first year did not have a continuous double auction prediction market. The comparison does not cover this period. Poll data from the first year were used to derive optimal parameters for poll aggregation algorithms. Experiment 1 and

Experiment 2 report results from the second and third year of the tournament, which ran from June 19, 2012 to April 10, 2013, and from August 1, 2013, to May 10, 2014, respectively.

Forecasters were recruited from professional societies' email lists, blogs, research centers, alumni associations, and personal connections. Entry into the tournament required a bachelor's degree or higher as well the completion of psychological and political knowledge tests, which took approximately two hours to compete. Participants were an average of 36 years old and mostly male (83%). Almost two-thirds (64%) had some postgraduate training.

3.1. Experimental Design

In this study, more than 1,600 individuals participated across 114 questions, amounting to approximately 65,000 and 54,000 forecasts from independent and team polls, and 61,000 market orders. Participants received $150 payment if they participated in at least 30 questions during the season.

We compared the accuracy of forecasts produced by two types of prediction polls (teams and individuals) and a prediction market (CDA) by randomly assigning individuals to the three experimental conditions.[4] There were no significant differences at baseline across the three experimental conditions. Although the number of forecasters assigned to prediction markets and prediction polls was about 700, we focus only on forecasters who made at least one forecast or market order. These samples ranged from 535 to 595. Participants assigned to prediction markets were less likely to submit at least one prediction than those in prediction polls (76% versus 81% and 85%, respectively). Attrition rates were 11%, 14%, and 18% for prediction markets, team, and independent prediction polls.[5] Table 1 shows demographic and retention data.

3.2. Elicitation Methods

In prediction markets, each contract traded at prices between $0 and $1 and resolved at $1 if the event occurred, $0 otherwise. Each participant started the season with $100 in cash and received additional $225 in increments of $5 over the course of the season. Price history was public information, as was the order book, which displayed the six highest bids and the six lowest ask prices. Forecasters had access to a dashboard, which showed their portfolio holdings, including profits or losses per question, as well as cash available to spend. The leaderboard featured the top 50 forecasters in terms of total net worth. The aggregate probability forecasts assessed were based on the last price as of midnight Pacific time.[6]

Table 1. Participants Across Elicitation Methods

	Prediction markets	Prediction polls	
		Teams	Individuals
Number with at least one order or one forecast	535	565	595
Gender (% male)	84	83	82
Age	35.9 (12.3)	35.1 (10.9)	35.8 (11.6)
Education (% with advanced degrees)	63	65	62
Proportion with at least one order or forecast (%)	76	81	85
Attrition rate (%)	11	14	18

In team and independent prediction polls, participants provided probability forecasts. Accuracy was measured using the Brier score, a proper scoring rule (Brier 1950). Scores varied from 0 (best) to 2 (worst). Forecasters in independent prediction polls worked independently. After a question resolved, performance was assessed as the mean daily Brier score.[7] Scores were averaged across questions, and the 50 forecasters with the lowest Brier scores were featured on a leaderboard.

In team polls, forecasters received a 30-minute training module about how to build a high-functioning team and hold each other accountable. Teams also had access to collaborative tools allowing them to write comments, use team mail to send personal messages, and receive information about the forecasts of their teammates. Accuracy scores for team members appeared on a within-team leaderboard. A separate leaderboard compared performance across teams. Team consensus was not required for submitting a forecast. Each team member submitted their own predictions and team scores were based on median forecaster's Brier score for each forecasting question.

The following analogies are useful in portraying the information structure of markets and polls. Prediction market traders form a mesh network, in which traders exchange information by placing orders. A hub-and-spokes analogy is most apt for describing independent polls: each individual makes solo predictions but is unable to exchange information with others. These independent predictions are aggregated statistically. Team-based polls resemble an archipelago, where each team is an island: forecasters can freely communicate within, but not across teams.

3.3. Questions and Scoring

Participants in Experiment 1 submitted probability estimates or market orders for 114 geopolitical questions using a Web interface. Questions were released throughout the forecasting season in small batches. Question duration varied between 2 and 410 days, with a median of 104 days. Forecasters were encouraged to update their beliefs as often as they wished until the question was resolved.

All questions had to pass the clairvoyance test: the resolution criteria had to be specified clearly enough to produce a definitive answer to the question. Another desirable question property was the 10/90 rule: at time of launch, events in question should be deemed more than 10%, but less 90% likely.

Results are based on aggregate forecasts. Probability forecasts and market orders were collected at the same time each day. Accuracy of both methods was assessed using the average daily Brier score, similar to the one used for individual prediction polling forecasters, except that aggregate forecasts were available every day each question was open, so carrying forward old forecasts was unnecessary.

3.4. Aggregation Algorithms in Prediction Polls

As discussed above, a critical component of poll-based prediction is the aggregation of individual predictions into a single forecast. The Good Judgment team used a weighted average:

$$\bar{p}_{t,k,l} = \left(\frac{1}{\sum_{t,i}\{d_{t,i,l} \times w_{t,i,l}\}}\right) \times \sum_{t,i}\{d_{t,i,l} \times w_{t,i,l}\} \times p_{t,i,k,l}, \qquad (1)$$

where $\bar{p}_{t,k,l}$ is the weighted probability aggregate across forecasters i, that depends on time t, question l, and outcome k. The decay value, $d_{t,i,l}$, was set to 1 for recent forecasts, and 0 otherwise. The decay parameter was based on 20% of total forecasters. For example, in a sample of 500 forecasters attempting a given question, 20% decay means that the most recently updated forecasts for 100 forecasters would be included.

The weight of a forecast, w, depended on time t, forecaster i, and question l. We used performance weighting with two variables: prior accuracy and frequency of belief updating. Weights for past accuracy were based on an individual's mean Brier score on all closed questions at that time. Raw Brier scores varied widely across questions: the average participant scored 0.02 on the "easiest" question and 1.5 on the "hardest." To account for this, we standardized scores within questions. Past accuracy weights were not set until the first 10 questions in a season had closed. Weights for belief-updating frequency were based on an individual's number of forecasts for the current question up to this point.

The overall weight was the product of weights for accuracy and frequency updating. Prior accuracy was the cumulative mean standardized Brier score for all resolved questions the forecaster attempted. Values were rescaled to the range [0.1 to 1]; the least accurate forecaster would receive 10% of the weight of the most accurate forecaster. The same rescaling was applied to the number of times a forecaster updated his or her probability estimate. Variables were combined as follows:

$$w_{t,i,l} = c^{\gamma}_{t,i} \times f^{\delta}_{t,i,l}, \qquad (2)$$

where the accuracy score, c, calculated at time t and forecaster i, was raised to the exponent γ. The frequency score, f, calculated at time t for forecaster i on question l, was raised to the exponent δ. Weights were exponents, and higher exponents produced weights with greater variation.

Finally, we recalibrated the aggregate forecasts as follows:

$$\hat{p}_{t,k,l} = \frac{\bar{p}^{a}_{t,k,l}}{\bar{p}^{a}_{t,k,l} + (1-\bar{p}_{t,k,l})^{a}}, \qquad (3)$$

where $\bar{p}$ is the original, nontransformed probability from the Equation (1), $\hat{p}$ is the transformed probability estimate, and a is the recalibration parameter. When $a = 1$, the aggregate is unchanged. At $a > 1$, aggregates are more extreme, at $a < 1$, aggregates are less extreme.

All parameters were set sequentially, starting with temporal decay, prior performance weighting, belief-updating weighting, and ending with recalibration, with the goal of minimizing error in the resulting aggregated forecasts. At each step, we used elastic net regularization (Zou and Hastie 2005) to minimize the overfitting. This method penalizes entry into the model and the magnitude of each parameter therein, which increases the chances that the parameters obtained at each step capture global versus local optimal weights.

4. Experiment 1: Results

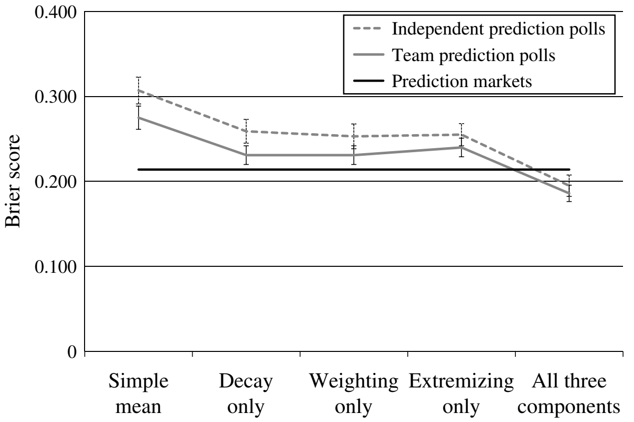
4.1. Relative Accuracy

We assessed the accuracy of methods using paired t-tests with questions as the unit of analysis. Prediction markets significantly outperformed the simple means of prediction polls. Mean Brier scores for team and independent polls were 0.274 and 0.307, respectively, significantly higher than the prediction market's mean Brier score of 0.214 (paired $t(113) = 4.45$ and $t(113) = 5.92$, $p < 0.001$ for each). Relative to team and independent polls, prediction markets were 22% and 30% more accurate. The results, shown in Figure 1, support Hypothesis 1.

4.1.1. Aggregation Parameters. Hypothesis 2 states that prediction poll forecasts that are combined statistically to reflect temporal decay, differential weighting, and recalibration are more accurate than prediction markets. To test this hypothesis, we estimated the parameters of the aggregation rule by fitting the algorithm to an independent set of year 1 data and applying it out of sample to year 2 data.

Figure 1 shows how each component of the aggregation rule decreased errors relative to the mean. With all three components, team prediction polls were significantly more accurate than the prediction market, yielding Brier scores that were 13% lower (mean difference = 0.027, paired $t(113) = 2.69$, $p = 0.008$). This result supports Hypothesis 2. Forecasts from

Figure 1. Mean Brier Scores for Independent Prediction Polls, Team Prediction Polls, and Prediction Markets with 114 Questions



Note. Errors bars denote one standard error of the difference between each polling method and the prediction market.

independent polls that were aggregated statistically outperformed prediction market prices by 9%, but the differences did not reach statistical significance ($t(113) = 1.51$, $p = 0.136$).

To summarize, results supported Hypothesis 1: prediction markets were more accurate than prediction polls when poll forecasts were combined using a simple mean. In addition, Hypothesis 2 was partially supported. Forecasts from team prediction polls combined using the full algorithm outperformed the prediction market. Furthermore, the accuracy of independent prediction polls and prediction markets was not reliably different. Prediction polls outperformed prediction markets when forecasters share information and aggregation was conducted with attention to timing, engagement, previous accuracy, and extremeness.

Next, we examine the stability and generalizability of the results by examining effects of (a) parameters used in the aggregation of forecasts from prediction polls, (b) different scoring rules, and (c) different measures of accuracy. Finally, we assess how relative accuracy varies with question duration and the number of active forecasters across questions.

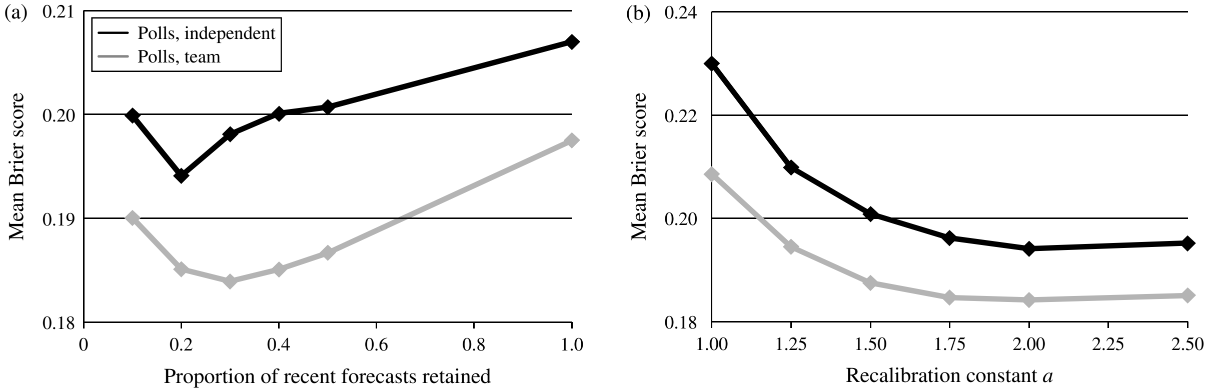
4.2. Sensitivity to Aggregation Parameters

The better performance of prediction polls was based on parameters derived from a prior tournament year (e.g., out-of-sample data), but it is possible that slight changes in parameter settings would have led to different results. We focus on two parameters: decay and recalibration. Figure 2(a) shows the effects of temporal decay parameters. Using a percentage of the most recent forecasts (20%), we obtain more accurate aggregates, and the exact proportion had a limited influence on the results.

Figure 2(b) shows effects of recalibration parameters. No recalibration, (denoted as $a = 1$) resulted in lower accuracy relative to recalibration parameters that pushed the aggregate forecasts toward the extremes ($a > 1$). However, a wide range of values for a produced comparable levels of accuracy. For example, the team recalibration parameter was $a = 1.5$. Results would have improved if $a = 2$, but the improvement in accuracy would have been less than 2%. To summarize, within limits, changes in both the parameters had limited impact on the accuracy of prediction polls. The advantage of team over independent polls was not a statistical artifact of our selection of aggregation parameters for the team forecasts.

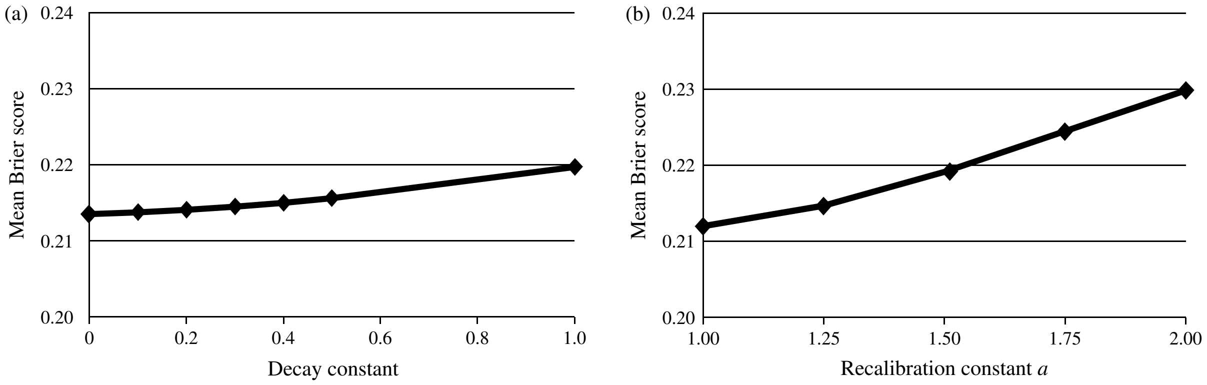
If statistical aggregation improved the accuracy of prediction polls, should statistical aggregation be applied to data from prediction markets? If market prices suffered from excess volatility, accuracy might increase with an exponential smoothing over the end-of-day prices. A decay function with an exponent of 0 means that old prices get no weight, which is the same as using the last price, whereas a decay parameter of 1 means that the last five end-of-day prices are equally

Figure 2. Aggregate Performance for Independent and Team-Based Prediction Polls, Varying Temporal Decay (a) and Recalibration (b) Parameters



Note. All other parameters remain at optimized levels.

Figure 3. Performance for Prediction Markets for Varying Temporal Decay (a) and Recalibration (b) Parameters



weighted. As shown in Figure 3(a), placing no weight on past prices achieved the highest accuracy.

Several studies have demonstrated that prediction markets suffer from the favorite long-shot bias (Page and Clemen 2012, Snowberg and Wolfers 2010). Rothschild (2009) showed that when market prices are recalibrated (pushed toward the extremes), accuracy improved. We applied the recalibration to market prices ($a = 1.5$). As we can see in Figure 3(b), extremizing market prices made accuracy worse, increasing Brier scores. Recalibration of prices away from the extremes also worsened accuracy. In summary, the last price yielded the most accurate probability forecast in the prediction markets.[8]

4.3. Alternative Scoring Rules

Do prediction markets outperform team prediction polls under different scoring rules? We examine three additional rules: logarithmic, spherical, and absolute distance. When a method assigns a value of f_c to the correct answer of a binary question, these rules are defined as follows:

$$\textit{Brier Scoring Rule} = 2 \times [1 - f_c]^2, \quad \text{from 2 (worst) to 0 (best)}$$

$$\textit{Logarithmic Scoring Rule} = \ln(f_c), \quad \text{from} - \infty \text{ (worst) to 0 (best)}$$

$$\textit{Spherical Scoring Rule} = f_c/[f_c^2 + (1 - f_c)^2]^{1/2}, \quad \text{from 0 (worst) to 1 (best)}$$

$$\textit{Absolute Distance Rule} = 2 \times |1 - f_c|, \quad \text{from 2 (worst) to 0 (best).}$$

All rules except the absolute distance rule are strictly proper, meaning that forecasters' optimal strategy is to report their true beliefs. Improvement of team prediction polls over prediction markets was 11% with the logarithmic rule, 2% with the spherical rule, and 6% with the absolute distance rule. Forecasts from team prediction polls were more accurate than prediction market prices with all scoring rules ($p < 0.05$). The independent prediction poll outperformed the prediction market with the absolute distance rule only. See Table 2.

4.4. Calibration, Discrimination, and AUC

Brier scores can be decomposed into three components: variability, calibration, and discrimination (Murphy 1973). Variability is independent of skill and simply reflects the base rate for events in the environment, so it will not be discussed further. Calibration refers to the ability to make subjective forecasts that, in the long run, coincide with the objective base rates of events. Calibration error would be zero if forecasts exactly matched event base rates. Probability predictions are said to be "well calibrated" if average confidence and the percentage of correct forecasts are equal. Before recalibration, team and independent prediction polls had calibration errors of 0.029 and 0.034 because of underconfidence. The prediction market produced lower calibration error, 0.009. Recalibration corrected the underconfidence in aggregate poll forecasts and improved the performance of prediction polls. This was evident from the decrease in Brier scores resulting from recalibration, as shown in Table 3. Figure 4 shows how recalibration decreased calibration errors for team and independent prediction polls.[9]

Table 2. Relative Performance as Judged by Four Scoring Rules

		Statistical aggregation	
Scoring rule	Prediction market	Team prediction poll	Independent prediction poll
Brier	0.21 (0.31)	0.19 (0.27)**	0.19 (0.31)
Log	−0.34 (0.41)	−0.30 (0.34)**	−0.31 (0.41)
Spherical	0.88 (0.18)	0.90 (0.16)*	0.89 (0.18)
Absolute distance	0.45 (0.40)	0.42 (0.36)*	0.40 (0.39)**

Note. Means (SD) across 114 questions.
$^{*}p < 0.05$; $^{**}p < 0.01$; paired samples t-test.

Table 3. Calibration and Discrimination for Statistically Aggregated Polls and Markets Based on Brier Score Decomposition

Method	Prediction market	Team prediction poll	Independent prediction poll
Calibration	0.009	0.010	0.008
Discrimination	0.355	0.384	0.375
Variability	0.585	0.585	0.585

Note. Recalibration based on year 1 was applied to year 2 prediction poll forecasts.

Discrimination refers to the tendency to place high-probability predictions on events that occur, and low estimates on those that do not. The discrimination values for statistically aggregated prediction polls were higher than those for the prediction market (see Table 3). Discrimination was better for both independent and team-based prediction polls with recalibration, relative to markets.

Area under the receiver operating characteristic curve (A-ROC) is a nonparametric measure of discrimination derived from signal detection theory (Swets 1996). The ROC curve plots the probability of a hit (true positive) against that of a false alarm (false positive). Perfect resolution means all of the probability mass is under the curve. ROC curves for prediction markets, as well as algorithmically aggregated independent and team polls are shown in Figure 5, grouped by early, middle, and late thirds of each question. In the first third of the questions, prediction markets yielded worse ROC scores than prediction polls. Markets and prediction polls were tied in the middle and late stages. The resolution advantage of team and independent prediction polls was stronger at the start of the questions, when there was maximum uncertainty about the outcome.

In summary, team and independent prediction polls either outperformed or tied prediction markets using scoring rules (logarithmic, spherical score, and absolute distance), alternative accuracy measures (discrimination and area under the ROC), and separate periods of the forecasting window (early, middle, and late). Next, we ask whether team prediction polls' performance edge varies with properties of the forecasting questions.

4.5. Duration of Forecasting Questions[10]

Does the relative performance of markets and polls depend upon the duration of the forecasting question? The favorite long-shot bias in prediction markets has been viewed as a rational response to liquidity constraints that occurs when questions are open for long periods of time. If payoffs do not materialize in a timely fashion, individuals may not wish to place expensive long-term bets on events, especially on favorites, because investments have opportunity costs (Page and Clemen 2012).

To examine whether prediction polls outperformed prediction markets on shorter and longer questions, we performed a median split based on the number of days questions were open (median duration was 105 days). We calculated daily Brier scores for each method on each question, averaged scores across questions, and rescaled question duration to the 0%–100% scale, where 0% denotes the first day and 100% denotes the last day of a question. Figure 6 shows the results of this exercise.

We used a mixed-effects linear model to determine whether daily Brier scores of prediction markets and team prediction polls varied by question duration, accounting for clustering of errors within questions and for the number of days until the question was resolved. We detected a significant interaction ($t = 4.32$, $p < 0.001$) between method (market versus team poll) and question duration (shorter versus longer questions). Prediction markets performed on par with team polls on shorter questions, but produced larger errors on longer questions. This pattern, evident in Figure 6, is consistent with Page and Clemen's (2012) discussion of the favorite long-shot bias, except the current results concern overall accuracy rather than calibration.

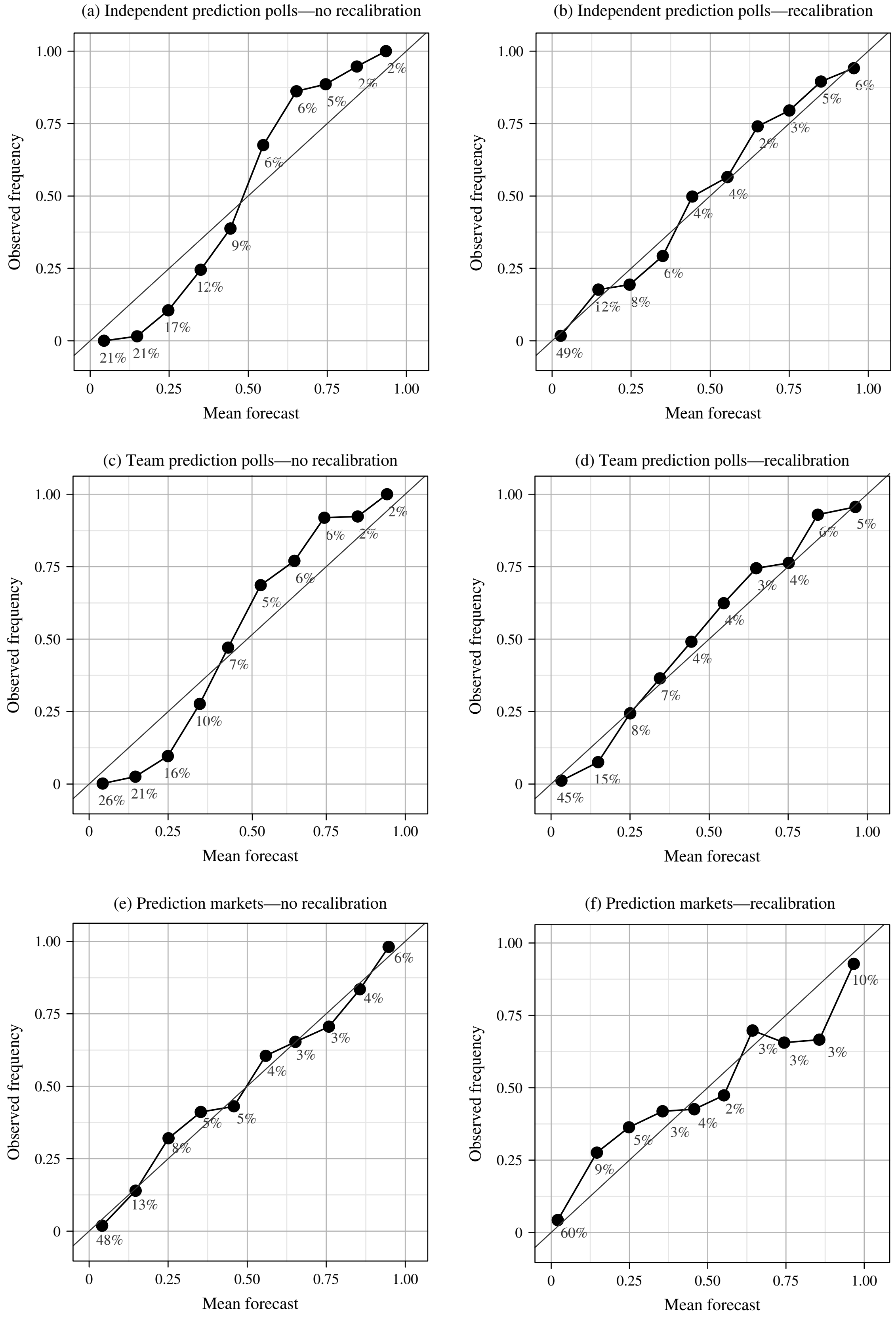
4.6. Forecasting Activity

Better performance in team prediction polls might have been due to greater forecaster activity relative to prediction markets. We examine two measures of activity—number of questions attempted, and engagement within a question. Engagement was defined differently across methods. In prediction polls, we used frequency of updating, and in the prediction market, we used number of orders.

As shown in Table 4, forecasters in the prediction market attempted fewer questions than those in prediction polls. Averages were 34, 52, and 66 in the market, the team poll, and independent poll, respectively. However, the average number of orders placed per question in the market was larger than the average number of forecasts made (over persons and questions) in the polls.[11]

Mellers et al. (2015a) found that, in prediction polls, frequency of updating was a better predictor of accuracy than the number of forecasting questions attempted. We replicated these results. In team and independent prediction polls, belief updating correlated with average standardized Brier scores ($r = -0.18$, $t(1{,}158) = 6.12$, $p < 0.001$), whereas number of questions attempted was not significant. A similar pattern held in prediction markets. The average number of orders placed per question was the best behavioral predictor of total earnings ($r = 0.55$, $t(534) = 20.05$, $p < 0.001$), whereas the number of questions attempted

Figure 4. Calibration Plots for Independent Prediction Polls, Before (a) and After (b) Recalibration; Team Prediction Poll Forecasts, Before (c) and After (d) recalibration; Prediction Markets, Before (e) and After (f) Recalibration ($a = 1.5$)



(a) Independent prediction polls—no recalibration

(b) Independent prediction polls—recalibration

(c) Team prediction polls—no recalibration

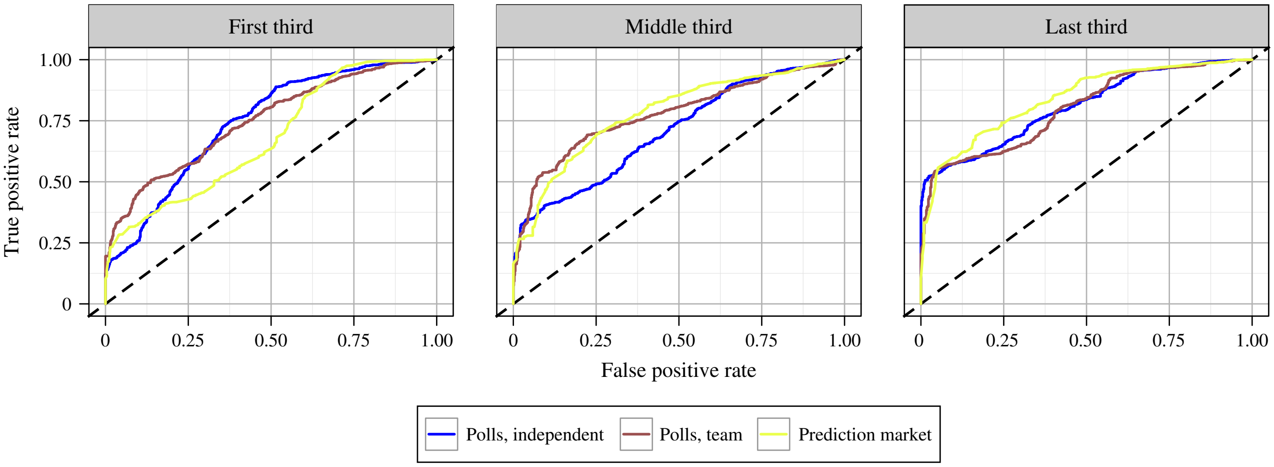
(d) Team prediction polls—recalibration

(e) Prediction markets—no recalibration

(f) Prediction markets—recalibration

Notes. Horizontal axis is divided in 10 bins, each spanning 10% on the probability scale. Percentages in chart denote the proportion of forecasts that fall in each bin.

Figure 5. (Color online) ROC Curves for Forecasts From First Third, Middle Third, and Last Third of Days When Each Question Was Open (Shown with Statistically Aggregated Polls and Prediction Markets)

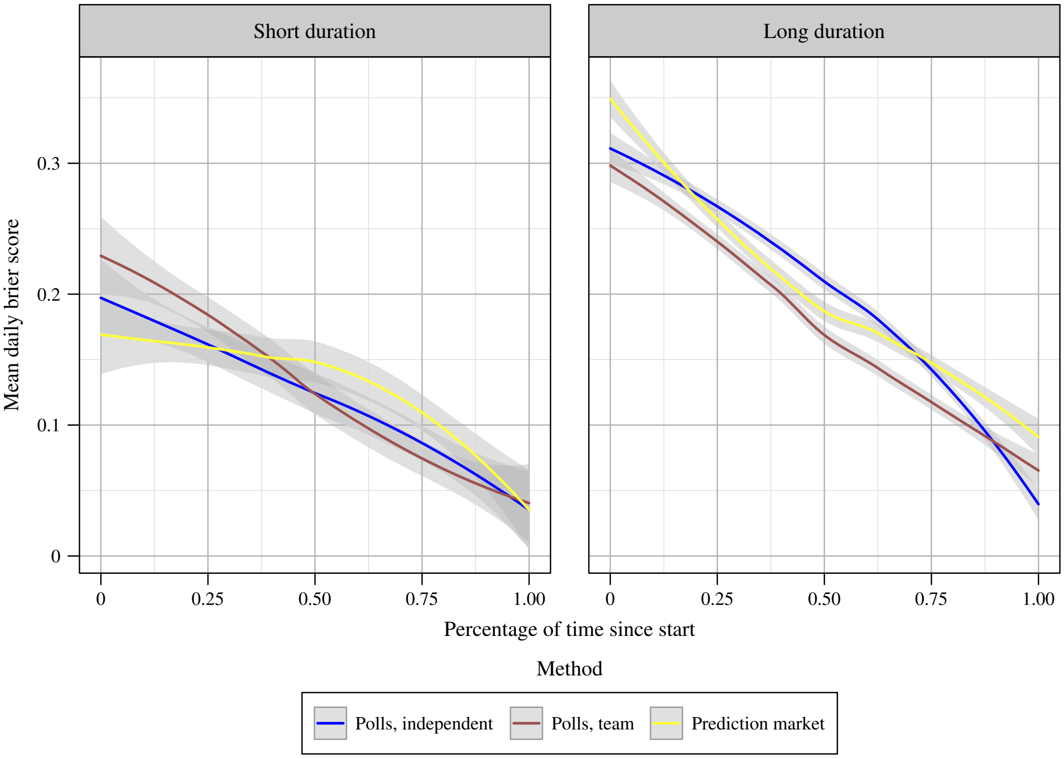


was a weaker predictor of earnings ($r = 0.20$, $t(534) = 4.80$, $p < 0.001$).

The focus on mean activity levels belies an important distinction between markets and polls. Prediction market orders were skewed across market participants, as evidenced by the higher standard deviation for orders per question. More than 50% of all orders were placed by the most active 5% of prediction market participants. In contrast, fewer than 30% of all forecasts were made by the most active 5% of prediction poll participants.

Simple forecast counts offer an incomplete picture of relative influence, due to the differential weighting of individuals in polls and the varying transaction sizes in markets. To add a more refined perspective, we calculated the relative weight of each forecaster on each question across all days the question was open, accounting for decay, accuracy, and updating weights. We then summed the weights across questions to obtain a measure of overall influence per forecaster throughout the season. The 20 highest-weighted forecasters in their condition accounted

Figure 6. (Color online) Performance Over Time for Statistically Aggregated Polls and Prediction Markets When Questions Lasted Less (Left) and More (Right) Than the Median, 105 Days



Note. Local regression (Loess) smoothed curves are added to illustrate differences over time.

Table 4. Activity Patterns Across Elicitation Methods

Activity metric	Prediction market	Team prediction poll	Independent prediction poll
Number of questions attempted	34 (26)	52 (42)	66 (45)
Number of orders/forecasts per forecaster per question	1.9 (2.9)	1.8 (1.5)	1.6 (1.8)
Number of orders/forecasts per forecaster per year	63 (106)	88 (107)	102 (134)
Number of forecasters per question	136 (66)	228 (85)	299 (85)
Number of orders/forecasts per question	260 (201)	473 (218)	491 (221)

Note. Cell entries are means; standard deviations shown in parentheses.

for 27% and 34% of the weights for team and independent polls, respectively.

We performed a parallel calculation in prediction markets, except we used total transaction values as a marker of influence. For example, a "buy" order for three shares at \$0.40 would have a total cost of \$1.20, and the matching sell order will cost \$0.60 per share or \$1.80 in total. We summed up transaction values for each person within a question, then calculated the proportion of all dollars spent by each participant across all questions. The top 20 traders accounted for 33% of the overall transaction values. To the extent that weights and transaction values can be compared as measures of relative influence across platforms, we can say that highly active forecasters tended to exercise more influence in markets than in team polls. Independent polls and prediction markets were generally similar.

The last row of Table 4 shows that, on average, more forecasts/orders were made per question in prediction polls than in prediction markets. Is the greater accuracy of team prediction polls driven by larger sample sizes? In the independent prediction polls conditions, we can estimate the impact of sample size by randomly selecting forecasters and removing them from the aggregate. These analyses showed that the market and poll performance was similarly sensitive to smaller samples. See the online appendix.

We now proceed to Experiment 2, which took place in the year following Experiment 1. This study provides a robustness check for Experiment 1 by applying the similar elicitation and aggregation methods to a new set of questions.

5. Experiment 2: Methods

More than 1,900 individuals participated between August 1, 2013, and May 10, 2014. One hundred and forty seven questions were asked. Participants received a \$250 payment conditional on forecasting in at least 30 questions over the course of the season.

Experiment 2 is based on data from the third year of the ACE tournament, directly following Experiment 1, which covered the second year. The experimental design in Experiment 2 followed the Experiment 1 convention: new forecasters were randomly assigned to the three experimental conditions; continuing forecasters remained in their original conditions (e.g., prediction market forecasters were reassigned to a market).

Experiment 2 differed from Experiment 1 in five ways. First, the number of participants assigned to teams prediction polls was larger than the number of forecasters in independent prediction polls ($N = 839$ and 550 in polls versus prediction markets with $N = 539$). Second, all market participants traded in one larger prediction market, rather than two smaller markets, as was the case in Experiment 1. Third, all team participants received forecasting training, whereas half of independent poll and prediction market participants received training. Fourth, half of the teams had a team coach to help facilitate discussion, whereas none of the independent forecasters and market forecasters had a coach. Coaching had no impact on accuracy. Fifth, only U.S. citizens were allowed in the prediction market, whereas approximately 22% of team and 24% of independent poll participants were non-U.S. citizens.

To address these experimental design challenges, we excluded all non-U.S. citizens from independent and team prediction polls. After this, we were left with 456 and 605 independent and team participants, respectively. The sample included 44,000 predictions in independent polls, 61,000 predictions in team polls, and 71,000 orders in prediction markets. Of course, exclusion of non-U.S. citizens from teams was not a complete solution. Although removing team members' probability was straightforward, we could not eliminate the information they shared with their U.S. teammates. The proportion of non-U.S. citizens in each team ranged from 9% to 42%, with a mean of 24%. There was no correlation between those proportions and team accuracy. Thus, there was no correlational evidence having more non-U.S. citizens in a team was associated with higher accuracy, but such a relationship cannot be ruled out. Despite these nonexperimental assignment features, we believe Experiment 2 provides additional evidence about the relative performance of methods.

6. Experiment 2: Results

The mean Brier score across all 147 questions for the prediction market was 0.24 (SD = 0.34). Simple means of team polls and independent polls were 0.22 (SD = 0.22) and 0.31 (SD = 0.20), respectively. The prediction market was more accurate than the simple mean of independent poll forecasts (paired $t(146) = 3.48$, $p < 0.01$), but no different from the simple mean

Table 5. Forecasting Accuracy for Experiment 2 Using Four Scoring Rules

		Statistical aggregation	
Scoring rule	Prediction market	Team prediction poll	Independent prediction poll
Brier	0.24 (0.33)	0.18 (0.29)**	0.25 (0.33)
Log	−0.40 (0.47)	−0.30 (0.44)**	−0.41 (0.53)
Spherical	0.86 (0.19)	0.90 (0.16)**	0.86 (0.19)
Absolute distance	0.49 (0.40)	0.36 (0.37)**	0.43 (0.42)*

Note. Means (SD) shown for 147 questions.
$^{*}p < 0.05$; $^{**}p < 0.01$; paired samples t-test.

of team poll forecasts. Thus, results were partially consistent with Hypothesis 1; prediction markets outperformed simple means of independent polls, but not simple means of team polls.

When we added temporal decay, differential weighting, and recalibration to the aggregation algorithm, team prediction polls significantly outperformed prediction markets for all four scoring rules, shown in Table 5. Independent polls were tied with markets, except for polls' significant outperformance according to the absolute distance rule.

7. Discussion and Conclusions

In Experiment 1, a randomized experiment comparing prediction markets versus competitive, dynamic prediction polls, simple aggregations of prediction polls were less accurate than market prices. The prediction market performed well with no additional benefit from statistical aggregation. The last price of the market was the single most accurate estimate that could be extracted from market data. Statistical aggregation improved accuracy in prediction polls, helping team polls to significantly outperform markets. This advantage held up across different scoring rules and accuracy measures. Team prediction polls were better than prediction markets on longer questions and at the early stages of the questions, the periods of greatest uncertainty. Prediction markets were more accurate than polls at the start of short-duration questions. The advantage of team polls over markets extended to Experiment 2.

7.1. What Gives Prediction Polls Their Edge?

The stronger performance of prediction polls can be attributed to several factors. First, to perform well in a prediction poll, one needs substantive knowledge and skill in converting insights into probability judgments. In prediction markets, one needs both timely, relevant information, and an understanding of the strategic aspects of market trading. The need for strategic sophistication in markets may have served as a barrier to high performance for otherwise knowledgeable forecasters.

Second, in settings where questions are posed and answered sequentially, information about prior activity and performance can be used to amplify the voices of the most careful and accurate forecasters in prediction polls and dampen the clamor of inaccurate ones. Prediction polls can use differential weighting based on prior performance. In prediction markets, individuals were weighted based on order size in the short run and accumulated wealth in the long run. Our results suggest that the weighting mechanisms used in prediction polls are at least as effective as those inherent to markets.

Third, crowd aggregates of probability judgments are usually underconfident, even when individual forecasts are not (Ranjan and Gneiting 2008). When such patterns persist over time, recalibration tends to improve aggregate accuracy. This problem was effectively addressed with the recalibration formula we used in both Experiment 1 and Experiment 2.

Fourth, the incentives for sharing information are better in team prediction polls than in markets. Continuous double auction markets are a zero sum game: one trader's gain is another's loss. By contrast, team prediction polls are structured so that people have incentives to discuss their beliefs and rationales with team members, even if they think someone else knows more. Teams are not limited to prediction polls; intrateam collaboration could, in principle, be incorporated into a prediction market.

Fifth, the advantage of polls is especially great when the number of active forecasters is small. A prediction poll of one person may be useful, but there is no market with a single trader. Even with a handful of people, the simple polling averages provide reasonable results (Mannes et al. 2014), while double auction prediction markets suffer from liquidity problems. Scoring rule markets (e.g., LMSR) are designed to solve the small market problem, but they require some interventions from tournament designers to set a liquidity parameter. Scoring rule markets are more similar to polls in that they do not operate as a zero-sum game. These markets may be more conducive to information sharing among forecasters. For settings with large numbers of forecasters per question, Hanson (2003) suggests that prediction polls may "suffer from a thick market problem, namely, how to produce a single consensus estimate when different people give differing estimates." (p. 108). We show that appropriate statistical aggregations can solve this problem.

In sum, team prediction polls created a mix of intrateam cooperation and interteam competition. This mixed environment may be better than a purely competitive one if individuals share information with each other and pool their knowledge. Team prediction polls may have a cognitive advantage via information sharing. Team prediction polls may have a motivational advantage if team members encourage each other to update their forecasts regularly. More complex forms of both cooperation and competition may increase accuracy relative to the purely competitive market environment. Ostrovsky (2012) discusses information pooling challenges in markets populated by strategic, partially informed traders. These challenges can be addressed by introducing incentives for cooperation, such as those used in team-based polls. In addition, teaming has been shown to improve problem solving in probability and reasoning tasks, including in auctions (Maciejovsky et al. 2013).

7.2. Trade-offs

There are pros and cons to prediction polls and prediction markets that go beyond pure accuracy. Prediction markets aggregate forecasts instantly and automatically. The instant aggregation and price availability to all market participants is a great way to promote transparency. Markets provide a decentralized method for information discovery. On the other hand, prediction polls perform well even when all participants provide independent estimates, with no crowd feedback, and are aggregated statistically. This feature may be especially useful in settings where managers need to collect predictions but have reasons to avoid sharing aggregate forecasts with participants.

Prediction polls require a statistical aggregation rule to more effectively pool multiple opinions. The need for devising statistical aggregation rules increases the responsibilities of the manager running the prediction polls. However, the components of effective statistical algorithms are fairly general, and parameter values for temporal decay and recalibration are relatively insensitive to misspecification. Individual weighting of forecasters is done after tournament designers achieve sufficient knowledge about participants. Based on our experience, “seeding” polls with 20–25 initial questions should be sufficient to derive stable aggregation parameters. Aggregation could easily be built into prediction polling software to provide a more seamless, off-the-shelf application.

7.3. Future Directions

Although we present the first long-term, experimental comparison of prediction polls and prediction markets, our results are by no means the final word. The tournament paid participants for participation rather than accuracy. If payments had been based on performance, the results could have differed. Studies of prediction markets have examined the effects of real performance-based payments relative to hypothetical-based payments, and the results are mixed. Servan-Schreiber et al. (2004) found no effects of real monetary incentives in prediction markets, but Rosenbloom and Notz (2006) showed that real financial incentives produced better forecasts. Camerer and Hogarth (1999) argued that real monetary incentives tend to decrease response variance. Finally, Hertwig and Ortmann (2001) found that performance incentives tend to induce decisions that more closely approximate rationality. Thus it is reasonable to expect that accuracy in both prediction markets and prediction polls with proper scoring rules will improve with the introduction of monetary incentives. However, it is less clear if the accuracy benefits of monetary incentives will be greater in markets or in polls.

One controversial feature of our design was the use of a leaderboard that displayed not only absolute earnings and proper scores, but also ranks. The distinction between absolute scores and ranks has been discussed in the expert elicitation and macroeconomic forecasting literature. Laster et al. (1999) discuss a model, and supporting evidence, that awards top forecasters for deliberately deviating from the consensus. Lichtendahl and Winkler (2007) argue that rank-based incentives push probability forecasters toward the extremes, relative to a setting with proper scoring rules without rankings. Although risk seeking tends to push polling averages toward the extremes, we note that it has the opposite effect in markets: risk-loving traders prefer long-shot bets, which tends to push prices away from the extremes.

The predicted patterns of overconfidence in polls and underconfidence in markets are inconsistent with our empirical results. Prediction poll forecasts are slightly overconfident at the individual level (Moore et al. 2015), but underconfident at the aggregate level, as we show above. Prediction market prices show neither systematic underconfidence nor overconfidence. Prices did not exhibit the typical favorite long-shot bias, implying that ranking system did not lead to excessive risk taking. The lack of distortions may be reassuring to companies that run prediction markets and rely on leaderboards and reputational incentives (Cowgill and Zitzewitz 2015).

Another limitation to the generality of our results is that we tested only continuous double auction markets. We cannot generalize our results to other market structures, such as pari-mutuel or logarithmic market scoring rule (LMSR) markets. Future research should examine these possibilities.

7.4. Conclusions

We show that crowds of several hundred individuals can produce highly accurate predictions on a wide range of political and economic topics. A comparison of two crowd-based forecasting approaches demonstrated that prediction polls that used teams and statistical aggregation were more accurate than prediction markets. Our results show that it is possible to take the elements that make prediction markets work well, such as incentivized belief elicitation, effective aggregation, and performance weighting, and combine these elements to a crowdsourcing method that produces even more accurate aggregate beliefs.

Acknowledgments

The authors thank INFORMS, Association for Psychological Science, and Judgment and Decision Making conference participants for helpful comments on previous drafts. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. 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.

Endnotes

[1] In the linear polls, forecasts were aggregated with an arithmetic mean. In the logarithmic polls, forecasts were aggregated with a geometric mean.

[2] Using a similar data set, Servan-Schreiber et al. (2004) found prediction markets to be more accurate than survey-based forecasts (from ProbabilityFootball), but the difference in results was traced to the imputation of missing forecasts. Servan-Schreiber et al. (2004) imputed 50% probability values for missing forecasts, which resulted in higher errors. Chen et al. (2005) omitted missing observations.

[3] Much of the finance literature is concerned with infinite markets in which individual traders have no ability to influence market prices. We use markets with several hundred forecasters. Individual traders can impact prices, at least in the short run.

[4] Additional conditions were investigated but are not discussed here (see Mellers et al. 2014).

[5] Sample sizes reported here are higher than those reported in Mellers et al. (2014), because the current study used data from participants who submitted at least one forecast or market order, whereas Mellers et al. (2014) focused on participants who submitted enough forecasts to evaluate their individual performance. The less restrictive inclusion criteria in the current study resulted in higher attrition rate estimates.

[6] In the first three months of the tournament, participants were randomly assigned to four markets of approximately 125 individuals. Then two markets of approximately 250 individuals each were formed. Earnings and positions were carried over to the combined markets. See the online appendix.

[7] For questions with three or more ordered answer options, we applied an ordered version of the scoring rule, which provides partial credit (or penalty reduction) for placing high probability estimates on answer options closer to the correct answer.

[8] We examined different ways of combining estimates from the two prediction markets and found the simple mean of latest market prices to be effective. For example, we calculated bid-ask spreads for the two parallel markets and constructed a weighted mean price, placing more weight on the market with the lower bid-ask spread for each question-day combination but this weighting scheme did not improve accuracy. See the online appendix.

[9] The underconfidence of aggregate team poll of forecasts is not relevant to debates about group polarization, groupthink or related phenomena. It is mathematically possible that each individual team produces polarized, overconfident predictions, but averaging across teams produces forecasts that are underconfident.

[10] We also examined other potential moderators, such as question-level uncertainty ratings, close call index ratings, status quo versus change outcome classification, question type (binary, conditional, and multioption). None of these variables were significant moderators of relative performance.

[11] Multiple orders or forecasts on a question on a given day only counted as one order or forecast.

References

Armstrong S (2001) Combining forecasts. Armstrong JS, ed. *Principles of Forecasting* (Springer, New York), 417–439.

Atanasov P, Rescober P, Servan-Schreiber E, Mellers B, Tetlock P, Ungar L (2013) The marketcast method for aggregating prediction market forecasts. *Proc. 2013 Internat. Conf. Social Comput., Behavioral-Cultural Modeling, and Prediction* (Springer, New York).

Baron J, Mellers BA, Tetlock PE, Stone E, Ungar LH (2014) Two reasons to make aggregated probability forecasts more extreme. *Decision Anal.* 11(2):133–145.

Berg J, Forsythe R, Nelson F, Rietz T (2001) Results from a dozen years of election futures markets research. Plott CR, Smith VL, eds. *Handbook of Experimental Economics Results*, Vol. 1 (North-Holland, Amsterdam), 742–751.

Brier GW (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* 78(1):1–3.

Camerer CF, Hogarth RM (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework. *J. Risk Uncertainty* 19(1–3):7–42.

Chen Y, Chu C, Mullen T, Pennock D (2005) Information markets vs. opinion pools: An empirical comparison. *Proc. 6th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 58–67.

Cowgill B, Zitzewitz E (2015) Corporate prediction markets: Evidence from Google, Ford, and Firm X. *Rev. Econom. Stud.* 82(4): 1309–1341.

Fama EF (1970) Efficient capital markets: A review of theory and empirical work. *J. Finance* 25(2):383–417.

Forsythe R, Nelson F, Neumann GR, Wright J (1992) Anatomy of an experimental political stock market. *Amer. Econom. Rev.* 82(5):1142–1161.

Goel S, Reeves D, Watts D, Pennock D (2010) Prediction without markets. *Proc. 11th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 357–366.

Graefe A, Armstrong J (2011) Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task. *Internat. J. Forecasting* 27(1):183–195.

Gürkaynak RS, Wolfers J (2006) Macroeconomic derivatives: An initial analysis of market-based macro forecasts, uncertainty, and risk. NBER Working Paper 11929, National Bureau of Economic Research, Cambridge, MA.

Hanson R (2003) Combinatorial information market design. *Inform. Systems Frontiers* 5(1):107–119.

Hayek F (1945) The use of knowledge in society. *Amer. Econom. Rev.* 35(4):519–530.

Hertwig R, Ortmann A (2001) Experimental practices in economics: A methodological challenge for psychologists? *Behav. Brain Sci.* 24(3):383–403.

Lichtendahl KC Jr, Winkler RL (2007) Probability elicitation, scoring rules, and competition among forecasters. *Management Sci.* 53(11):1745–1755.

Laster D, Bennett P, Geoum IS (1999) Rational bias in macroeconomic forecasts. *Quart. J. Econom.* 114(1):293–318.

Maciejovsky B, Sutter M, Budescu D, Bernau P (2013) Teams make you smarter: How exposure to teams improves individual decisions in probability and reasoning tasks. *Management Sci.* 59(6):1255–1270.

Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. *J. Personality Soc. Psych.* 107(2):276–299.

Manski C (2006) Interpreting the predictions of prediction markets. *Econom. Lett.* 91(3):425–429.

Mellers B, Ungar L, Baron J, Ramos J, Gurcay B, Fincher K, Tetlock P (2014) Psychological strategies for winning a geopolitical forecasting tournament. *Psychol. Sci.* 25(5):1106–1115.

Mellers B, Stone E, Atanasov P, Rohrbaugh N, Metz SE, Ungar L, Bishop MM, Horowitz M, Merkle E, Tetlock P (2015a) The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *J. Experiment. Psychol.: Appl.* 21(1):1–14.

Mellers B, Stone E, Murray T, Minster A, Rohrbaugh N, Bishop M, Chen E, et al. (2015b) Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspect. Psychol. Sci.* 10(3):267–281.

Moore DA, Swift SA, Minster A, Mellers B, Ungar L, Tetlock P, Yang H, Tenney R (2015) Confidence calibration in a multiyear geopolitical forecasting competition. Working paper, University of California, Berkeley, Berkeley.

Murphy AH (1973) A new vector partition of the probability score. *J. Appl. Meteorology* 12(4):595–600.

Ostrovsky M (2012) Information aggregation in dynamic markets with strategic traders. *Econometrica* 80(6):2595–2647.

Page L, Clemen RT (2012) Do prediction markets produce well-calibrated probability forecasts? *Econom. J.* 123(568):491–513.

Pennock DM, Lawrence S, Nielsen FA, Giles CL (2001) Extracting collective probabilistic forecasts from Web games. *Proc. 7th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 174–183.

Plott C, Chen K (2002) Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem. Working paper, California Institute of Technology, Pasadena.

Ranjan R, Gneiting T (2008) Combining probability forecasts. Technical Report 543, University of Washington, Seattle.

Rieg R, Schoder R (2010) Forecasting accuracy: Comparing prediction markets and surveys—An experimental study. *J. Prediction Markets* 4(3):1–19.

Rosenbloom ES, Notz W (2006) Statistical tests of real-money versus play-money prediction markets. *Electronic Markets* 16(1):63–69.

Rothschild D (2009) Forecasting elections: Comparing prediction markets, polls, and their biases. *Public Opinion Quart.* 73(5): 895–916.

Rothschild D, Wolfers J (2010) Forecasting elections: Voter intentions versus expectations. Working paper, National Institute of Economic Research, Cambridge, MA.

Satopää VA, Jensen ST, Mellers BA, Tetlock PE, Ungar LH (2014a) Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *Ann. Appl. Statist.* 8(2): 1256–1280.

Satopää VA, Baron J, Foster DP, Mellers BA, Tetlock PE, Ungar LH (2014b) Combining multiple probability predictions using a simple logit model. *Internat. J. Forecasting* 30(2):344–356.

Servan-Schreiber E (2007) About competitive forecasting. White paper, NewsFutures, Paris.

Servan-Schreiber E (2012) Prediction markets: Trading uncertainty for collective wisdom. Landemore H, Elster J, eds. *Collective Wisdom: Principles and Mechanisms* (Cambridge University Press, New York).

Servan-Schreiber E, Wolfers J, Pennock DM, Galebach B (2004) Prediction markets: Does money matter? *Electronic Markets* 14(3):243–251.

Sherden WA (1998) *The Fortune Sellers: The Big Business of Buying and Selling Predictions* (John Wiley & Sons, Hoboken, NJ).

Silver N (2012) *The Signal and the Noise: Why So Many Predictions Fail—But Some Don't* (Penguin, New York).

Snowberg E, Wolfers J (2010) Explaining the favorite-long shot bias: Is it risk-love or misperceptions? *J. Political Econom.* 118(4): 723–746.

Sonnemann U, Camerer C, Fox C, Langer T (2013) How psychological framing affects economic market prices in the lab and field. *Proc. Natl. Acad. Sci. USA* 110(29):11779–11784.

Spann M, Skiera B (2003) Internet-based virtual stock markets for business forecasting. *Management Sci.* 49(10):1310–1326.

Surowiecki J (2005) *The Wisdom of Crowds* (Anchor Books, New York).

Swets J (1996) *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers* (Lawrence Erlbaum Associates, Mahwah, NJ).

Tetlock PE (2005) *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton University Press, Princeton, NJ).

Tetlock PE, Mellers BA, Rohrbaugh N, Chen E (2014) Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions Psych. Sci.* 23(4):290–295.

Wolfers J (2009) Prediction markets: The collective knowledge of market participants. *CFA Institute Conf. Proc. Quart.* 26(2):37–44.

Wolfers J, Zitzewitz E (2006) Interpreting prediction market prices as probabilities. NBER Working Paper 12200, National Bureau of Economic Research, Cambridge, MA.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J. Royal Statist. Soc.* 67(2):301–320.