Judgment and Decision Making, Vol. 14, No. 2, March 2019, pp. 135–147

Are markets more accurate than polls? The surprising informational value of "just asking"

Jason Dana* Pavel Atanasov† Philip Tetlock† Barbara Mellers†

Abstract

Psychologists typically measure beliefs and preferences using self-reports, whereas economists are much more likely to infer them from behavior. Prediction markets appear to be a victory for the economic approach, having yielded more accurate probability estimates than opinion polls or experts for a wide variety of events, all without ever asking for self-reported beliefs. We conduct the most direct comparison to date of prediction markets to simple self-reports using a within-subject design. Our participants traded on the likelihood of geopolitical events. Each time they placed a trade, they first had to report their belief that the event would occur on a 0–100 scale. When previously validated aggregation algorithms were applied to self-reported beliefs, they were at least as accurate as prediction-market prices in predicting a wide range of geopolitical events. Furthermore, the combination of approaches was significantly more accurate than prediction-market prices alone, indicating that self-reports contained information that the market did not efficiently aggregate. Combining measurement techniques across behavioral and social sciences may have greater benefits than previously thought.

Keywords: prediction, forecast, judgment, prediction markets, self-reports, surveys

1 Introduction

Behavioral and social scientists have long disagreed over how best to measure mental states. While psychologists clearly value behavioral measures, they quite often measure beliefs and preferences by simply asking people to self-report them on a numerical scale. And while economists place value on people's judgments, they tend to place greater value on inferring preferences and beliefs from behavior. For example, if a person claims that the United States is on the verge of an economic collapse or that a climate disaster is imminent, an economist might look at that person's investment portfolio or disaster preparedness to *reveal* whether that person really believes these statements. Indeed, the unwillingness of economists to rely on survey questions has been called an important divide with other behavioral scientists (Bertrand & Mullainathan, 2001).

Perhaps the most impressive demonstration of the power of using revealed beliefs is the resounding success of prediction markets. Prediction markets create contracts that pay a fixed amount if an event occurs, and then allow people to trade on the contract by submitting buying or selling prices in a manner similar to the stock market. The price at which the contract trades at a given time can be taken to be the market's collective probability estimate of the event occurring. For example, suppose the event to be predicted was the winner of the 2016 US presidential election, and that a contract paid $100 if Hillary Clinton won. If the contract last traded at $60 – that is, someone just purchased the contract from someone else for $60 — one could use that price as a likelihood prediction of Clinton winning of 60%. In other words, if a risk-neutral market is valuing a risky $100 contract at $60, it implies that the expected value of the contract is $60 and thus that it will pay out with probability .6.

Using prediction market prices in this manner has yielded impressively accurate predictions for a wide array of outcomes, such as the winners of elections or sporting events, typically exceeding the accuracy of "just asking" methods such as opinion polls or expert forecasts (see Wolfers & Zitzewitz, 2004; Ray, 2006, for reviews). When market participants have some intrinsic interest in trying to predict results, even markets with modest incentives or no incentives have been shown to be effective. As examples, small markets using academics as participants predict which behavioral sci-

All raw data are publicly available via the Open Science Framework and can be accessed at https://dataverse.harvard.edu/dataverse/gjp.

B. Mellers, J. Dana, P. Atanasov and P. Tetlock developed the study concept. J. Dana and P. Atanasov developed relevant tests and performed all data analyses. J. Dana, P. Atanasov, and B. Mellers wrote the manuscript with critical input from P. Tetlock. All authors approved the final version of the manuscript for submission.

We thank Phillip Rescober for data assistance and Paul Tetlock and Joi Ito for helpful discussions.

This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center (DoI/NBC) Contract No. D11PC20061. 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.

*Yale University. Email: jason.dana@yale.edu.

†Department of Psychology, University of Pennsylvania

https://doi.org/10.1017/S1930297500003375 Published online by Cambridge University Press

ence experiments will successfully replicate (e.g., Camerer et al., 2018) and "play money" markets in which participants play for prestige can be as accurate as real-money markets (Pennock et al, 2001; Servan-Schreiber et al., 2004). Because these probability forecasts are obtained without ever asking anyone to self-report their beliefs, the success of prediction markets appears to be a victory of the economic approach and a repudiation of relying on self-reports.

The classic explanation for why prediction markets are so successful is that they are efficient mechanisms for integrating information useful to making predictions. To see why, suppose that someone had information that suggested an event was much more likely to occur than the current market price suggested. That person would now have the incentive to buy the contract because its expected value would greatly exceed its cost. The balance of such beliefs would eventually push the price up. Others might have pieces of information that suggest the event is unlikely, motivating them to sell and putting downward pressure on the price. In the end, the market price will tend to reflect the balance of information that participants have. Indeed, when traders try to engage in market manipulation, buying and selling with the intent of changing the price to provide misinformation, their attempts usually fail and the market can become even more informative due to the incentives for traders to act on their true beliefs (Hanson, Oprea & Porter, 2006). In theory, there should be no information in their self-reports that is not already reflected in the market price. The success of prediction markets appears to support that theory.

It is difficult, however, to draw clean inferences from in-the-wild comparisons of prediction markets and self-reports for several reasons. Traders in markets are necessarily exposed to information from others in the market, such as historical prices, the last price at which shares traded, and the current buy and sell orders. In this way, traders may be working with more information than poll respondents. Further, prediction markets aggregate opinions in a unique way. The market price is the point at which optimistic and pessimistic opinions "cross". The market price is thus a marginal opinion that is not simply an average or a vote count (Forsythe et al., 1992). It could be that the magic of prediction markets lies largely in superior aggregation methods rather than superior quality or informativeness of responses. Finally, selection issues could be serious when comparing market participants with poll respondents. Participation in prediction markets is nearly always self-selected and people who choose to trade might be different by having more intrinsic interest or knowledge or better analytic skills. We are unaware of a comparison between surveys and prediction markets that addresses all of these problems, so it is not even clear that in such a comparison, "just asking" would be inferior.

We had a unique opportunity to compare methods in an experiment that addressed all of these issues and put both approaches on a level playing field. During the IARPA Aggregative Contingent Estimation (ACE) tournament (Mellers et al., 2014; Atanasov et al., 2017), our team, the Good Judgment Project, randomly assigned participants to take part in a prediction market. Each time participants wanted to place an order, they were first asked to report their beliefs that an event would occur on a 0 to 100 probability scale. We then aggregated these self-reports in a pre-determined fashion using best practices gleaned from earlier years of the tournament, such as extremizing the aggregate, weighting recent opinions more heavily than older ones, and weighting forecasters with a good track record more heavily. When aggregated this way, simple self-reports were at least as accurate as market prices for predicting a variety of geopolitical events.

Perhaps more importantly, a combination of prices and self-reports was significantly better than prices alone, indicating that self-reports contained incrementally useful information that market prices alone did not capture. One could wonder whether "just asking" was good in our study or whether our methods of aggregation were good. But proper aggregation can have only limited benefits if the responses do not contain information. Our results suggest that self-reports not only contained useful information, but information that was not efficiently captured by the market price.

Prior research that is perhaps most similar to ours was done by Goel et al. (2010), who compared the accuracy of prediction markets with opinion polls and simple statistical models. Across thousands of American football games, betting markets were found to have only a tiny edge over opinion polls. Across multiple domains, very simple statistical models approached the accuracy of prediction markets, suggesting diminishing returns to information; nearly all predictive power was captured by 2 or 3 parameters. These studies, however, were not experimental. Different people participated in the polls and markets, raising the inferential problems noted above. Further, it is unclear how general the comparisons between markets and polls were, because they involved American football predictions. Fans of American football are inundated with statistical information and betting lines, and therefore their opinions might be highly correlated with betting markets, which would naturally lead to similar accuracy. Here, we employ 113 different geopolitical events, usually lasting months and ranging from typical prediction market domains (e.g., who will win a national election) to the exotic (e.g., will Kenneth Bae leave North Korea or will construction begin on the Lamu pipeline before a given date). The full list of prediction questions is included in the Appendix.

Our study also bears similarities to research on "the wisdom of the crowd within" (Vul & Pashler, 2008) or "dialectical bootstrapping" (Herzog & Hertwig, 2009), in which averaging multiple judgments from the same person exceeds the accuracy of the individual judgments themselves. We find a similar increase in accuracy when we combine aggre-

https://doi.org/10.1017/S1930297500003375 Published online by Cambridge University Press

gated self-reports and prediction market prices, which are themselves determined by individual bids. While bids and self-reported probability beliefs can be seen as two different measures, they are highly correlated and thus can be seen as two "draws" from the same personal generating process.

2 Method

Five hundred thirty-five volunteers participating in the third year of the ACE tournament sponsored by the Intelligence Advanced Research Project Activity (IARPA) were randomly assigned to participate in a prediction market (see Atanasov, et al., 2017). In the prediction market, participants could buy or sell shares of events at prices between 0 and 100, where 0 represented the closing value of the shares if the event did not occur and 100 represented the value of the shares if the event did occur. Before they could complete their buy or sell orders, participants also had to report their belief in the probability of the event occurring on a 0 to 100 scale.

Our experimental design thus permits a better comparison between self-reports and prediction markets because the participants saw the same information — the last trading price and the bid and asking prices in the market — when making trades and self-reporting beliefs. This design also eliminates self-selection concerns because the market participants were randomly assigned to the prediction market and other conditions from a larger pool. Lastly, the same group of participants both made trades and judged probabilities.

2.1 Participants

We recruited forecasters into the larger participant pool from professional societies, research centers, alumni associations, science blogs, and word of mouth. Participation required a bachelor's degree or higher and completion of a battery of psychological and political tests that took an average of 2 hours. Participants were U.S. citizens and mostly males (80%); with an average age of 43. Over two thirds (70%) had postgraduate training.

2.2 Questions and measures

The dataset consisted of 113 forecasting questions with binary yes/no outcomes. Questions with more than two outcomes raised complications because participants were not asked to give a probability for each outcome, just the one they were betting on. Examples of questions included, "Will China seize control of the Second Thomas Shoal before 1 January 2014?" or "Will India and/or Brazil become a permanent member of the U.N. Security Council before 1 March 2015?" Questions remained open for an average of 102 days. A list of all questions is given in the Appendix and criteria for resolving questions can be found at https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/BPCDH5/L8WZEF.

For each question, participants saw the prices and number of shares requested for the six highest buy orders and the six lowest sell orders. They could bid or ask for shares at the price that they specified. Whenever participants entered an order, they stated their belief that the event would occur on a 0 (certain it will not occur) to 100 (certain it will occur) probability scale before their order could be confirmed. Participants were encouraged to return to the Web site and update their predictions at any time until the question closed. Throughout the year, participants could trade on any questions they wished until either the events were resolved or the trading year closed. Our data included 46,168 market orders and self-reported beliefs. Of those, we focus primarily on the 37,000 of those orders that were matched into trades, because those orders contributed directly to market prices.

2.3 Incentives

Participants who made trades and self-reported beliefs on at least 25 events throughout the year were paid for their participation with a $250 Amazon gift card. Participants competed for social rewards, including a place on the leaderboard and the chance to join an elite group of "superforecasters", but payments were not tied directly to their performance in the market.

2.4 Aggregating self-reported probability beliefs

The aggregate was formed using an algorithm (Atanasov et al., 2017) whose parameters were determined using data from other forecasters in a prior year of the competition. The algorithm had three features. First, more recent self-reports were given priority over older ones because questions were open for some time and older self-reports become outdated as new information becomes available. Based on prior years, we used only the 20% most recent self-reports at any time. Second, greater weight was assigned to the beliefs of forecasters who had a track record of accuracy. For all questions that had resolved as of the date of the opinion, participants were scored for the accuracy of their self-reported beliefs. Their opinions were then given weights ranging from .1 to 1 for the worst to the best proportional to their Brier scores on past predictions, and those weights were raised to an exponent of 4, which was determined to minimize the Brier score of the aggregate error in previous years. Third, the aggregate was extremized toward 0 or 1 because measurement error pushes individual estimates toward the middle of a probability scale and because individual estimates neglect the information that is in the other estimates (see Baron et al., 2014). We used Baron et al.'s transformation formula

https://doi.org/10.1017/S1930297500003375 Published online by Cambridge University Press

for a probability aggregate p, $\frac{p^a}{p^a+(1-p)^a}$ with $a = 2$ based on prior year data. We used an elastic net technique to avoid overfitting. For a more detailed description of this aggregation procedure and the logic behind it, see Atanasov et al. (2017).

3 Results

3.1 Quality of self-reported beliefs and their relationship to market orders

Because participants were primarily recruited to participate in a prediction market and were not incentivized to give accurate self-reports of their beliefs, we first looked for signs that these judgments were taken seriously. Participants' order prices were generally consistent with their stated beliefs, with the two being similar, though not identical (r=0.66). This result suggests that the self-report question was taken seriously, but also that the two modes of answering could potentially yield different information, since there is still a substantial amount of unshared variance between the two.

Another way to explore whether the self-reported probability judgments were taken seriously is to use them to imply the expected profit margins of the participant's market orders. The expected profit margin is simply the participant's reported probability minus the order prices for buy orders, and order price minus the probability for sell orders. If we are to take self-reported probability judgments seriously, one would expect these profit margins to be positive. For example, if a forecaster placed a buy order at a price that was higher than her judged belief (e.g. \$65 and 45%), she expects to *lose* \$20 for each contract she buys. Only 16% of the orders had implied negative profit margins. Further investigation reveals that traders were less likely to transact at negative profits margins when entering new positions (13%) than when reversing existing positions (28%). This pattern suggests that some of these negative-profit trades simply reflect risk aversion on the part of participants who pay a premium to take profit on a trade that has already proven successful relative to the current market price. Generally, then, we conclude that stated beliefs at least pass the surface test of coherence.

Finally, we can assess the quality of self-reported beliefs by examining their relationship to trading success. All else equal, we would expect that participants whose reported probabilities proved more accurate would also have better trading success. We calculated total earnings from all closed questions after the market season closed. We also calculated the mean standardized *Brier scores* (the squared error between the forecast and the 0 or 1 outcome) for each self-reported belief on each resolved event and converted these scores to ranks so that 100% was the best and 0% was the worst. The rank conversion ensured that the distributions were well-behaved in the presence of outliers. For forecasters who made trades on at least 25 events (i.e., those who were eligible for payment for their participation), belief accuracy and earnings were highly correlated (r=0.55).

Table 1: Brier scores for prices and beliefs.

Brier Score	vs. Prices	
Prices	0.227 (0.31)	
Beliefs, full algorithm	0.210 (0.39)	t(113)=1.44, p=0.152
Prices & beliefs	0.210 (0.34)	t(113)=2.92, p=0.004

Note: Brier score comparison for 113 forecasting events, with probability beliefs aggregated using frequency weights, a discounted function over time, and an extremizing transformation.

3.2 Relative accuracy of market prices and self-reports

Our goal was to compare the accuracy of prices and self-reported beliefs. "Prices" are simply the last trading prices on a question at midnight, Pacific Standard Time, for a given day. "Beliefs" are the aggregated self-reported probability judgments.

We calculated the Brier score measure of accuracy for each method (Brier, 1950). For questions with binary outcomes, Brier scores range from 0 to 2, where 0 is best and 2 is worst. Suppose the prediction for a two-outcome event was that the event was 70% likely to occur (and 30% likely to not occur) and the event occurred. The Brier score would be calculated as $(.7-1)^2+(.3-0)^2$, which equals 0.18. On the other hand, if the event did not occur, the Brier score would be $(.7-0)^2+(.3-1)^2$, which equals 0.98. If a forecaster gave a probability of 50%, the Brier score would be 0.5. We averaged scores over days and questions for each method.

Table 1 compares Brier scores for Prices and Beliefs. A simple, unweighted average of self-reports yielded a mean Brier score of .283 relative to a mean Brier score of 0.227 for Prices. One might be tempted to conclude that "just asking" thus yielded inferior prediction market prices. We resist this conclusion for a few reasons. First, as noted above, simple averaging of probability estimates inappropriately represents the information that they contain (Baron et al., 2014). Second, a comparison of a simple average of self-reported beliefs with Prices is awkward because prediction markets not only elicit information, but they aggregate it in a way that is not a simple average and does not weight everyone equally, so it would not be clear which factor lends more to the relative success of prediction markets. Finally, in the simple average of self-reports, old answers that were based on stale information were weighted just as much as new ones. The question we address is whether self-reports somehow fail to

https://doi.org/10.1017/S1930297500003375 Published online by Cambridge University Press

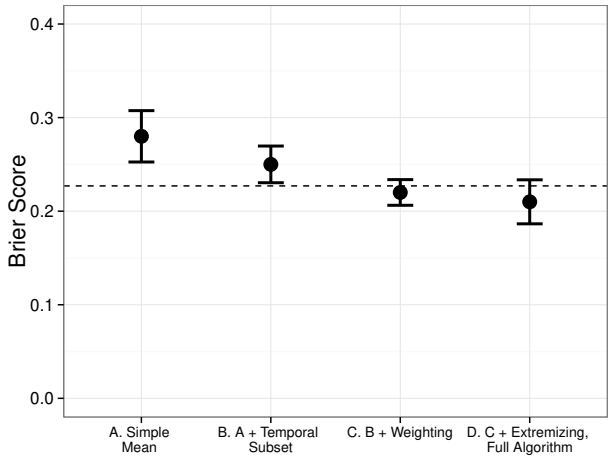

FIGURE 1: Brier scores for aggregated forecasts, last-price aggregates are shown as the horizontal solid line, the dotted-line shows Brier scores for aggregation algorithms of Beliefs, starting from simple mean, then adding temporal subsetting, past accuracy and update frequency to weights, and extremizing. Error bands denote two standard errors of the Brier scores difference between prices and aggregated beliefs.

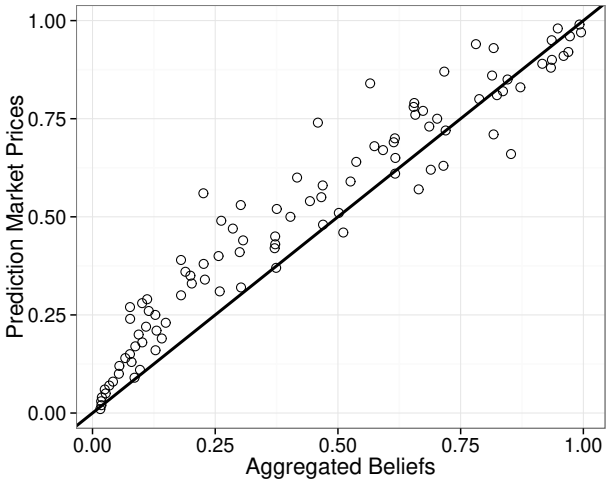

FIGURE 2: Prediction market prices versus aggregated beliefs. Aggregated belief probability values are averaged for every dollar increment in in market prices, from $1 to 99$, which correspond to probability values of 0.01 to 0.99. Aggregated beliefs tend to produce lower estimates than market prices, which is denoted positioning of points above the diagonal line. Prices are somewhat less extreme as well. (Note: While extremization of aggregated beliefs boosts accuracy, extremizing market prices does not reduce Brier scores.) (See Corrigendum.)

elicit quality information, and we report the accuracy of the simple average so that the reader may see the relative contribution to accuracy of the different factors involved in the aggregation algorithm.

Removing older information by keeping only the 20% most recent self-reports improved the accuracy of aggregated self-reports to a mean of 0.249. Adding weights based on the prior accuracy of the person giving the self-report further lowered mean Brier scores to 0.217. Finally, adding belief extremization reduced mean Brier scores to 0.210. Thus, Beliefs yielded a 12% reduction in Brier score over Prices, though this difference was not statistically significant (paired $t(112) = 1.44$, $p = .15$, within-subjects standardized $d = .14$). See Figure 1. In practical terms, this Brier score difference is modest, equivalent to assigning a probability of 66.3% to the correct answer for Prices and a probability of 67.6% to the correct answer for Beliefs.

Figure 2 plots the relationship between aggregated self-reports and market prices. For every dollar increment in in market prices from \$1 to \$99, we aggregated corresponding self-reported beliefs. Not surprisingly, the two are strongly related ($r = .95$). As denoted by the datapoints to the left of the diagonal below the median price and to the right above the median price, aggregated beliefs are more extreme. This is again unsurprising because the algorithm for aggregating beliefs involved an extremizing transformation. We note, however, that while extremizing modestly improved the Brier scores of Beliefs consistent with theory, it did not improve Prices.

We also computed the Brier score for the simple mean of Beliefs and Prices. This hybrid probability estimate yielded a significant improvement over Prices alone (paired $t(112) = 2.92$, $p = 0.004$, within-subjects standardized $d = .27$) and directionally outperformed Prices on 85% of the 113 forecasting questions.[1,2]

This result suggests that there was incrementally valuable information in self-reported beliefs that was not captured by market prices. If different groups had participated in the market and given self-reports, one could wonder if accuracy would be better still if the self-report group had participated in a prediction market and the prices from the two markets were averaged. In the present design, we can clearly infer that self-reports added informational value above and beyond the market.

[1]Although the combination had a similar Brier score to Beliefs, this comparison yielded more reliable differences, with smaller standard errors, thus the differences in t- and p-values.

[2]The direction of performance differences held when using mean absolute error rather than mean squared error (i.e. Brier scores) as a measure of accuracy. Both Beliefs and Belief-Price hybrid yielded better mean absolute errors that Prices alone, $t > 4.00$, $p < .001$ for both comparisons.

https://doi.org/10.1017/S1930297500003375 Published online by Cambridge University Press

TABLE 2: Brier score differences for prices and beliefs.

Intercept	–0.042 (0.019) *
Bid-ask spread	0.197 (0.009) *
Months to question resolution	0.014 (0.003) **
Observations	11,251
Questions	113

Note: Relative performance of Prices and aggregated Beliefs. Positive values denote worse performance for last price. Larger bid-ask spreads and longer months to resolution are associated with greater accuracy in Beliefs. Standard errors are in parentheses; * p < .05, ** p<.01.

3.3 Determinants of accuracy

When might self-reports be especially useful? When market volume is thin, the spread between the highest buying price and the lowest selling price is large, so people are less likely to trade, and the last market price is less informative. Large spreads between highest buying prices and lowest selling prices (bid-ask spreads) suggest low engagement in the forecasting question. In these cases, informed participants have fewer incentives to trade. In addition, sophisticated probability polls have been shown to outperform market prices when the resolution date of the forecasting question is in the distant future (Atanasov et al., 2017). Page and Clemen (2013) describe a related tendency for longer time until expiration to distort accuracy in the direction of a favorite-long shot bias: market participants are unwilling to lock in funds on relatively expensive bets on favorites when the question will take a long time to resolve, instead preferring small bets on long-shots, pushing market prices away from prices denoting extreme probabilities.

We examined the relationship between relative performance of the measure and the bid-ask spread using general estimating equations (GEE), which permit clustering of errors within questions. The criterion was the Brier score difference between Prices and Beliefs for any question on any given day, and predictor variables were bid-ask spreads (ranging from 0% to 100%), and number of months to question resolution. Positive regression coefficients mean that Prices performed worse than Beliefs.

Table 2 shows the results. The intercept was negative, meaning that Prices were more accurate than Beliefs on days immediately before question resolution and markets had minimal bid-ask spreads ($b = -0.042$, *Wald test* = 4.73, $p = 0.030$). Greater bid-ask spreads were associated with worse performance of Prices relative to Beliefs ($b = 0.197$, *Wald test* = 5.03, $p = .025$). Prices were also worse relative to Beliefs when the resolution of the question was far away into the future ($b = 0.014$, *Wald test* = 16.04, $p < .001$).

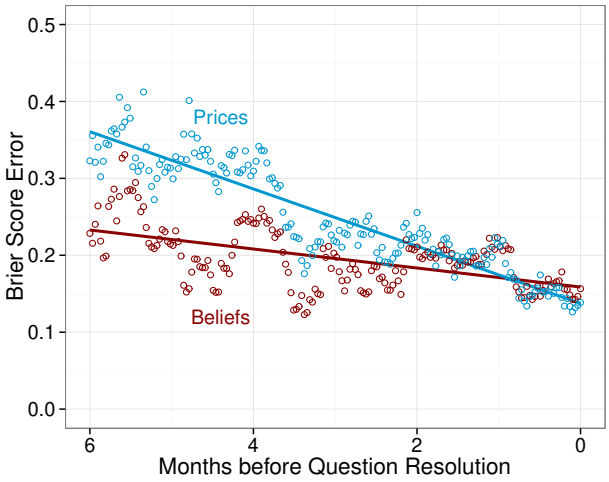

FIGURE 3: Note: Prices refers to the last market price on a given day as the probability estimate. Aggregated Beliefs are derived from forecasters' beliefs stated on the 0–100 scale, aggregated with a statistical algorithm (Atanasov, et al., 2017). Question sets varied over time; more questions were open closer to resolution. Lines depict ordinary least sequares model fits for each method.

Figure 3 shows Brier scores against months before question resolution. Each line shows the predicted effect of each method over time. Beliefs were better than last prices when the resolution of the question was months away, but prices were better right before the question resolved. As the regression in Table 2 shows, the direction and significance of this pattern persisted even when adjusting to market liquidity, as measured by bid-ask spreads – the lowest price of current sell orders minus the highest price of current buy orders.

3.4 Factors influencing relative accuracy

What are the reasons behind the market's relative inability to process information? We examine two possibilities. First, as suggested above, markets could be miscalibrated for far-off forecasts because participants simply do not want to lock up their trading funds. Second, market participants may be unable to find trading partners earlier on. That is, they may have placed a bid or an ask on a question, but nobody may have matched that bid or ask and thus, no trade occurred. Unmatched bids and asks do not affect market prices directly, but the participants placing those bids and asks may still have useful information to contribute when asked to self-report their beliefs.

To assess the first possibility, we used the Murphy & Winkler (1987) Brier score decomposition function to calculate calibration and discrimination for Prices vs. Beliefs, separately by first vs. second half of each question's duration.

https://doi.org/10.1017/S1930297500003375 Published online by Cambridge University Press

TABLE 3: Brier score decomposition.

	Brier score	Calibration error	Discrimination
First half of question duration			
Prices	0.271	0.021	0.249
Beliefs, full algorithm	0.232	0.014	0.283
Prices & beliefs	0.240	0.008	0.268
Second half of question duration			
Prices	0.185	0.019	0.334
Beliefs, full algorithm	0.191	0.031	0.341
Prices & beliefs	0.181	0.014	0.333

Note: Brier score comparison for 113 forecasting events, with probability beliefs aggregated using frequency weights, a discounted function over time, and an extremizing transformation.

In the calculation of these scores (as well as Brier scores above), each question is weighted equally, independent of its duration. The decomposition results are shown in Table 3 and illustrated in calibration plots in Figure 4. The higher accuracy of Beliefs in the early stages of questions is mostly due to superior discrimination (higher scores denote better performance) – the ability of self-reports to simply identify which events will occur and not. As shown in Figure 4, Panels A & C, neither method is perfectly calibrated early on in questions, as both methods tend to over-predicting events to occur relative to observed base rates.[3] For example, the fifth value in Panel C shows that among the forecasts signifying between 41%-50% probability of event occurrence, the events occurred in only about 25% of cases. In the second half of question duration, calibration improved for both methods, with self-reported beliefs showing signs of slight over-confidence, i.e., predictions more extreme than observed rates.

To examine the role of unmatched orders, we extracted and aggregated self-reports from matched orders and compared their accuracy to Beliefs calculated from all orders, matched or unmatched. We can thus compare the marginal value of self-reports associated with unmatched orders. We found that adding these reports did not improve the accuracy of Beliefs. Furthermore, the marginal benefit of self-reports associated with unmatched orders did not vary over time. Thus, neither technical explanation was empirically supported; it appears that self-reports indeed contained useful and unique information that the market did not capture, particularly when the event being predicted were farther into the future.

Activity in the markets was not evenly distributed. Of the 535 prediction market forecasters who placed at least one market order, the top 10 most active forecasters by number of orders accounted for approximately 33% of all market orders, while the top 50 accounted for 60%. In terms of share volume, the top 10 most active participants placed 44% of the shares ordered, while the top 50 accounted for 70% of ordered share volume. Still, less active forecasters had some influence over trading, at the very least by providing counterparties for trading against the most active members. Market designers generally do not see inequality of activity as a problem, but rather as a feature of the market structure: the most accurate participants tend to accumulate wealth funds over time, and are able to place more and larger orders, while market participants who find themselves on the wrong side of most bets tend to lose their wealth and thus the ability to participate in markets and influence market prices.

Finally, we consider the marginal benefits of eliciting beliefs in addition to market orders because collecting beliefs may be costly in some settings. How would much would self-reported beliefs add if we had limited ability to collect them? To answer this, we randomly selected date-question combinations and deleted the aggregated belief values. On those days, the hybrid prediction was based on prices alone. We then rescored the accuracy of the hybrid forecasts in subsamples where beliefs were available on 25%, 50%, 75% of date-question combinations. At the ends of the spectrum are cases where beliefs are available 0% of the time, thus only prices are used, and where beliefs are available 100% of the time. Both of these are reported above. In Table 4, we show median Brier scores across ten iterations of each

[3]This pattern of over-predicting events early on is not irrational from an individual perspective. Many of the questions have contingent stops. For example, the question asks if an event will occur before a certain date. If the event occurs, the question is resolved and contracts are immediately paid out. If the event does not occur by the deadline, traders have more time to exit their positions. Even if they do not exit, the losses would be realized later. So contract structure markets toward overpredicting events resulting in early closure. But it appears that this bias affects last price and aggregated beliefs approximately equally.

https://doi.org/10.1017/S1930297500003375 Published online by Cambridge University Press

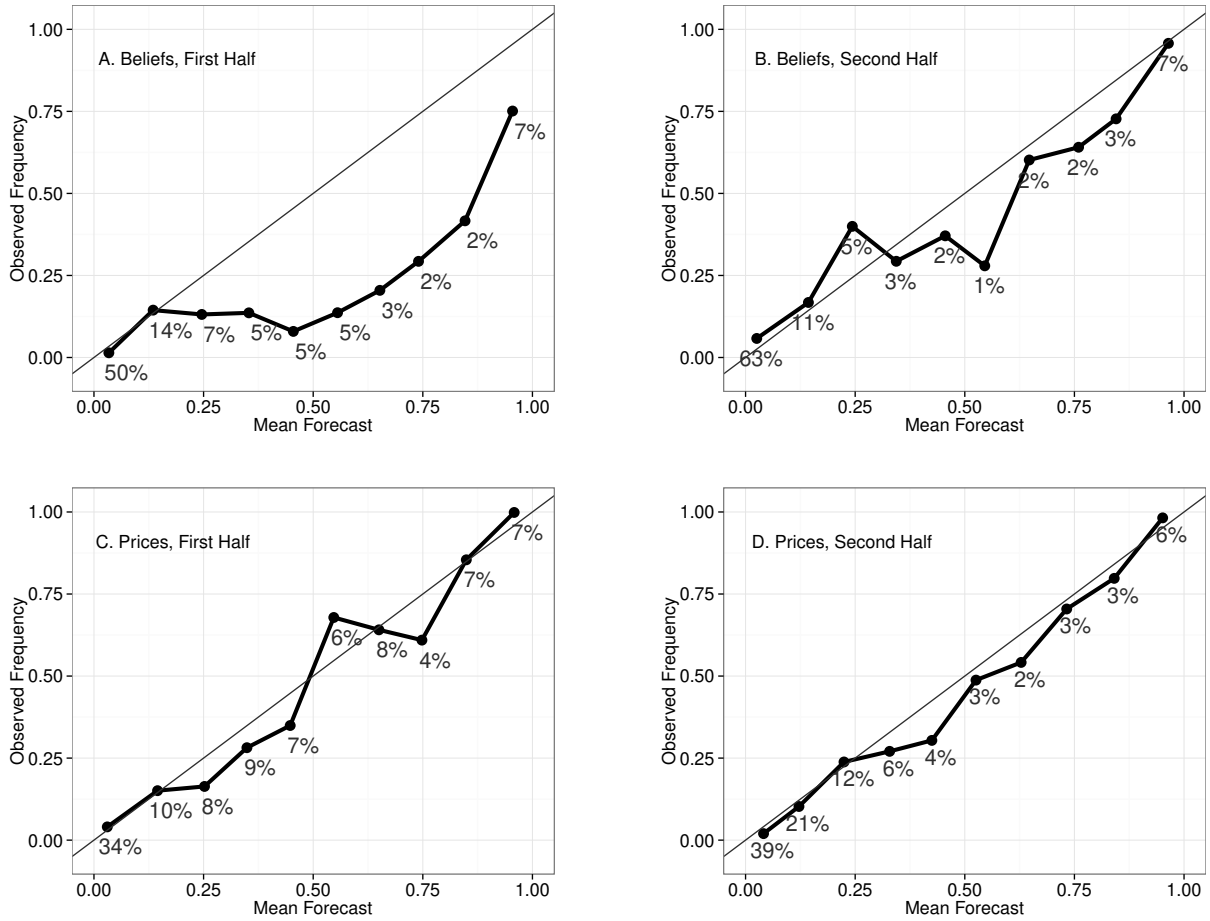

FIGURE 4: Calibration plots for Beliefs aggregated with full algorithm (A & B), vs. Last Prices (C & D), for first and second half of question duration. Forecasts are divided in 10 ordered bins (0%-10%, 11%-20%, etc.), and the mean forecast in each bin is plotted against the observed frequency of occurrence of predicted events. Points to the right of the 45-degree line denote overestimation of probability.

simulated level of belief availability. The relationship between aggregate belief availability and accuracy appears to be approximately linear. As the availability of aggregated beliefs is reduced, Brier scores increase.

We also examined if availability of beliefs at specific points is especially valuable at certain points of time. With the benefit of hindsight, we can surmise that beliefs would be especially useful on early days within a question, where beliefs tend to outperform prices (See Figure 3). To test this, we simulated the performance of the hybrid method if beliefs were only available for the first 30 days of each question's duration. If a question was open for more than 30 days, beliefs were available for the full period. In this case, beliefs were available for 30% of question-date combinations. Making beliefs available at the beginning of questions was particularly helpful in improving accuracy, with the hybrid method registering a Brier score of 0.218, slightly better than the simulated Brier score of 0.219, registered when beliefs were made available randomly, on a larger proportion (50%) of dates.

4 Discussion

Although it can be difficult to measure an attitude or belief better than asking people to simply self-report it, mistrust of self-reports in some areas of social science remains strong (Bertand & Mullainathan, 2001). An apparent triumph of measuring beliefs using behavioral methods has been the relative success of predictions markets. We used a design in which prediction market participants also provide self-reported beliefs every time they submitted an order. This design holds information constant across methods. For approximately 37,000 forecasts from 535 participants for 113 unique events, we found that self-reported beliefs were at

https://doi.org/10.1017/S1930297500003375 Published online by Cambridge University Press

TABLE 4: Median Brier scores across ten iterations of each simulated level of belief availability.

Availability of Beliefs	Brier score
100% of dates	0.210
75% of dates	0.214
50% of dates	0.219
25% of dates	0.223
0% of dates (i.e. prices only)	0.227
First 30 days only	0.218

Note: Brier scores across all questions for hybrid method combining prices and aggregated beliefs. Results are based on simulations in which beliefs are made unavailable for a random set of date-question combinations, or for dates beyond the first 30 days (last row).

least as informative as prediction market prices when beliefs were properly aggregated. Prediction markets do not simply average opinions or count votes. Similarly, our predetermined aggregation procedure gave more weight to recent opinions than older ones, weighted opinions by the prior success of people giving them, and extremized the aggregate, as is appropriate when people express probabilistic beliefs (Baron et al., 2014). Notably and consistent with theory, extremizing helped the aggregate, but did not help market prices.

Perhaps more importantly, self-reports appeared to provide unique information that the market mechanism did not integrate. Self-reports were correlated with market bids, as would be expected if people took the question seriously. But there was a substantial amount of non-shared variance in reports and bids. When the methods were combined, accuracy was significantly improved over prices alone, indicating that distinctively useful information was contained in the self-reports. While we do not know *what* precise information was communicated through belief reports, we learned *when* belief reports were most helpful in improving accuracy: at times with relatively low trading activity, high-bid ask spreads, and when questions are expected to resolve within months, rather than days or weeks.

Simple self-reports have perhaps been under-rated in their capacity to yield probability forecasts. But some caveats are in order. Participants in the Good Judgment Project often display high levels of motivation and put lots of time into the task. They are also highly educated compared to the population at large (70% had some postgraduate training). They also had no incentive to misrepresent their beliefs for the questions we asked them. In situations where participants are not motivated or might want to misreport their beliefs, self-reports may be deficient. But there are situations where participants are reasonably motivated and can be expected to take the question seriously, such as organizations making internal predictions. Several firms have attempted to use prediction markets for internal applications (see, e.g., Cowgill & Zitzewitz, 2015; Healy et al., 2010). When information is complex and the number of participants limited, alternative methods like iterated polls have been found to outperform markets (Healy et al., 2010). Our results lend further and broader support to the idea that firms could be better off polling properly incentivized participants and putting more effort into aggregating opinions properly.

Skeptics might question whether our prediction markets had enough active traders. Lack of liquidity can reduce the accuracy of prediction markets because trades occur too infrequently and do not reflect the most current state of information. The opposite can be true, however, as more liquid real-money prediction markets have been shown to elicit trades from naïve individuals during informative time periods such that prices can deviate further from fundamentals (Tetlock, 2008). We have found that low-liquidity markets can produce accurate predictions, and vice versa. That said, self-reported beliefs were especially helpful in our study when markets lacked liquidity, suggesting that indeed prices were inefficient at aggregating new information in these situations. Another concern is that participants were not trading for monetary rewards. They did, however, compete for social rewards, including a place on the leaderboard and the chance to join an elite group of "superforecasters." Moreover, this concern is probably over-stated; both public prediction markets and those within organizations have succeeded without financial incentives (Servan-Schreiber et al. 2014; Cowgill & Zitzewitz, 2015; Plott & Chen, 2002).

Our results show that self-reports can capture information that markets do not. There is a complementarity in the two approaches, and one can obtain more accurate forecasts by using both. Many psychologists have been surprised by how accurately market prices have estimated probabilities in a variety of domains. Many economists may now be surprised by how informative it is to "just ask."

References

Atanasov, P., Rescober, P., Stone, E., Servan-Schreiber, E., Tetlock, P., Ungar, L. & Mellers, B. (2017) Distilling the wisdom of crowds with prediction markets and prediction polls? *Management Science, 63*, 691–706.

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

Camerer, C., Dreber, A., Holzmeister, F. Ho, T., Huber, J., Johannesson, M., Kirchler, M. et al. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Be-*

https://doi.org/10.1017/S1930297500003375 Published online by Cambridge University Press

havior, 2, 637–644. https://doi.org/10.1038/s41562-018-0399-z.

Cowgill, B. & Zitzewitz, E. (2015). Corporate prediction markets: Evidence from Google, Ford, and firm X. *The Review of Economic Studies, 82*(4), 1309–1341.

Forsythe, R., Nelson, F., Neumann, G. R., & Wright, J. (1992). Anatomy of an experimental political stock market. *The American Economic Review, 82*, 1142–1161.

Goel, S., Reeves, D., Watts, D., & Pennock, D. (2010). Prediction without markets. *Proceedings of the 11th ACM conference on Electronic Commerce*, 357–366.

Hanson, R., Oprea, R., & Porter, D. (2006). Information aggregation and manipulation in an experimental market. *Journal of Economic Behavior and Organization, 60*, 449–459.

Healy, P., Linardi, S., Lowery, J., & Ledyard, J. (2010). Prediction markets: Alternative mechanisms for complex environments with few traders. *Management Science, 56*, 1977-1996.

Herzog, S., & Hertwig, R. (2009). The wisdom of many in one mind: improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*, 231–237.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S., Moore, D., Atanasov, P., Swift, S., Murray, T., & Tetlock, P. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115.

Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, *115*(7), 1330–1338.

Page, L. & Clemen, R. T. (2013). Do prediction markets produce well-calibrated probabilityforecasts?, *The Economic Journal*, *123*, 491–513.

Pennock, D.M., Lawrence, S., Giles, C.L., & Nielsen, F.A. (2001). The real power of artificial markets. *Science, 291*, 987–988.

Plott, C. & Chen, K. (2002) Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem. California Institute of Technology Social Science Working Paper 1131. Retrieved from https://authors.library.caltech.edu/44358/1/wp1131.pdf.

Ray, R. (2006). Prediction markets and the financial "wisdom of crowds." *Journal of Behavioral Finance, 7*, 2–4.

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

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

Tetlock, P.C. (2008). Liquidity and prediction market efficiency. Available at SSRN: http://ssrn.com/abstract=929916.

Thurstone, L. L. (1959). *The measurement of values*. Oxford, England: University of Chicago Press.

Vul, E. & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*, 645–647.

Wolfers, J. & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, *18,* 107–126.

Wolfers, J. & Zitzewitz, E. (2006). Interpreting prediction market prices as probabilities (No. w12200). National Bureau of Economic Research.

Appendix: Events to be predicted in the study

1. Will Iran blockade the Strait of Hormuz before 1 January 2014?
2. Will the World Trade Organization (WTO) rule in favor of the rare earth metals complaint filed by the European Union against China before 31 December 2013?
3. Will either the French or Swiss inquiries find elevated levels of polonium in the remains of Yasser Arafat's body?
4. Will the Taliban and the Afghan government commence official peace talks before 1 September 2013?
5. Will Angela Merkel win the next election for Chancellor of Germany?
6. Will North Korea attempt launch of a multistage rocket between 7 January 2013 and 1 September 2013?
7. Will Russia maintain any military presence at the Tartus Naval Base in Syria as of 1 January 2014?
8. Will a foreign state or multinational coalition officially announce a no-fly zone over Syria before 1 January 2014?
9. Will the Syrian government commence official talks with Syrian opposition forces before 1 September 2013?
10. Will a significant North Korean military force violate the Military Demarcation Line (MDL) of the Korean Demilitarized Zone (DMZ) before 1 October 2013?
11. Will there be a significant lethal confrontation in the East China Sea region between Japan and China before 1 January 2014?
12. Will Turkey ratify a new constitution before 1 February 2014?
13. Will Uhuru Kenyatta be found guilty of any charges by the International Criminal Court before 1 September 2013?
14. Will China seize control of the Second Thomas Shoal before 1 January 2014?
15. Before 1 May 2014, will Myanmar officially announce that construction of the Myitsone Dam will resume?
16. Before 1 May 2014, will Chinese armed forces or maritime law enforcement forces attempt to interdict or make physical contact with at least one U.S. government naval vessel or airplane or Japanese government naval vessel or airplane that it claims is in its territorial waters or airspace?

17. Before 1 May 2014, will Iran abolish the office of President of the Islamic Republic?
18. Will six-party talks with North Korea resume before 1 January 2014?
19. Before 1 May 2014, will Nicolas Maduro vacate the office of President of Venezuela?
20. Before 1 January 2014, will the government of Bolivia invite the U.S. Agency for International Development (USAID) to resume work in Bolivia?
21. Before 1 January 2014, will the government of Afghanistan sign a Status of Forces Agreement (SOFA) permitting U.S. troops to remain in Afghanistan?
22. Will Libya complete elections for a Constitutional Commission before 1 October 2013?
23. Will India and/or Brazil become a permanent member of the U.N. Security Council before 1 March 2015?
24. Will Chad experience an onset of insurgency between October 2013 and March 2014?
25. Will China deploy any armed unmanned aerial vehicles (UAVs) over the territory of another country before 1 May 2014?
26. Will China sell at least one unmanned aerial vehicle (UAV) to any other country before 1 May 2014?
27. Will Guinea commence legislative elections before 1 October 2013?
28. Before 1 May 2014, will Joseph Kony be captured or incapacitated by a Ugandan, foreign or multinational military/law enforcement force?
29. Before 1 December 2013, will Egypt impose a constitutional ban on political parties based on religion?
30. Before 1 April 2014, will the International Atomic Energy Agency (IAEA) inspect the Parchin Military Complex?
31. Before 1 May 2014, will Iran test a ballistic missile with a reported range greater than 2,500 km?
32. Before 1 February 2014, will either India or Pakistan recall its High Commissioner from the other country?
33. Will Prince Khalifa bin Salman Al Khalifa be Prime Minister of Bahrain on 1 February 2014?
34. Will Syria attack Israel between 28 August 2013 and 31 December 2013?
35. Will Nawaz Sharif vacate the office of Prime Minister of Pakistan before 1 May 2014?
36. Before 1 March 2014, will Gazprom announce that it has unilaterally reduced natural-gas exports to Ukraine?
37. Will the Organization for the Prohibition of Chemical Weapons (OPCW) complete its initial on-site inspections of Syria's declared chemical weapons sites before 1 December 2013?
38. Before 1 March 2014, will North Korea conduct another successful nuclear detonation?
39. Before 1 May 2014, will any non-U.S. actor use, in a lethal confrontation, either a firearm containing a critical part made with 3D printing technology or a lethal explosive device containing a critical part made with 3D printing technology?
40. Between 25 September 2013 and 31 March 2014, will any members or alternate members of the 18th Central Committee of the Communist Party of China be arrested on charges of bribery, embezzlement, or abuse of power?
41. Before or during its next plenary meeting, will the Central Committee of the Communist Party of China announce that it plans to reform the hukou system nationwide by 2015?
42. Before 1 May 2014, will Russia sign an agreement with the de facto government of South Ossetia delineating the border between the two?
43. Will the M-PESA system have a failure that results in at least 100,000 subscribers losing all ability to send and receive money from their accounts for at least 48 hours before 31 December 2013?
44. Before 1 May 2014, will the government of Colombia and the FARC sign a formal peace agreement?
45. Before 1 May 2014, will any U.N. member state offer diplomatic recognition to the government of a new state on what is now territory of Syria, Turkey, or Iraq?
46. Will Venezuela experience an onset of domestic political crisis between December 2013 and April 2014?
47. Before 1 December 2013, will the government of Pakistan and Tehrik-i-Taliban Pakistan announce that they have agreed to engage in direct talks with one another?
48. Will the president of Brazil come to the United States for an official State Visit before 1 February 2014?
49. Before 1 May 2014, will construction begin on the Lamu oil pipeline?
50. Will the INC (India National Congress) win more seats than any other party in the Lok Sabha in the 2014 General Elections in India?
51. Before 1 April 2014, will the government of Syria and the Syrian Supreme Military Command announce that they have agreed to a cease-fire?
52. Will defense expenditures in Japan's initial draft budget for fiscal year 2014 exceed 1 percent of projected gross domestic product (GDP)?
53. Will the United Kingdom's Tehran embassy officially reopen before 31 December 2013?
54. Will Facebook and/or Twitter be available in China's Shanghai Free Trade Zone before 31 March 2014?
55. Before 1 May 2014, will Russia rescind its law barring US citizens from adopting Russian children?
56. Before 1 February 2014, will Iran officially announce that it has agreed to significantly limit its uranium enrichment process?
57. Before 1 May 2014, will the government of any country other than Armenia, Belarus, Kazakhstan, Kyrgyzstan, Russia or Tajikistan announce its intention to join the Eurasian Customs Union?
58. Before 1 April 2014, will one or more countries impose a new requirement on travelers to show proof of a polio

vaccination before entering the country?
59. Before 1 January 2014, will the Prime Minister of Japan visit the Yasukuni Shrine?
60. Will Russia file a formal World Trade Organization (WTO) anti-dumping dispute against the European Union (EU) before 31 March 2014?
61. Before 1 May 2014, will China arrest Wang Zheng on charges of incitement to subvert state power and/or subversion of state power and/or incite separatism?
62. Will the general elections in Guinea-Bissau commence on 16 March 2014 as planned?
63. Between 4 December 2013 and 1 March 2014, will the European Commission officially state that Italy is eligible for the investment clause?
64. Will South Korea and Japan sign a new military intelligence pact before 1 March 2014?
65. Will North Kosovo experience any election-related violence before 31 December 2013?
66. Before 1 March 2014, will the U.S. and E.U. officially announce that they have reached at least partial agreement on the terms of a Transatlantic Trade and Investment Partnership (TTIP)?
67. Before 1 March 2014, will the European Commission (EC) announce that Turkey is permitted to open a new chapter of accession negotiations?
68. Before 31 March 2014, will the Slovenian government officially announce that it will seek a loan from either the European Union bailout facilities or the IMF?
69. Before 1 May 2014, will General Abdel Fattah al-Sisi announce that he plans to stand as a candidate in Egypt's next presidential election?
70. Before 1 May 2014, will the U.S. and the European Union reach an agreement on a plan to protect individuals' data privacy?
71. Before 1 May 2014, will official representatives of the Syrian government and the Syrian opposition formally agree on a political plan for Syria?
72. Will the six-party talks with North Korea resume before 1 May 2014?
73. Before 1 March 2014, will the International Atomic Energy Agency (IAEA) announce that it has visited the Gchine uranium mine site in Iran?
74. Before 31 March 2014, will either Peru or India announce their intention to formally launch negotiations on a preferential trade agreement (PTA) with each other?
75. Will Israel release all of the 104 Palestinian prisoners from its jails before 1 May 2014?
76. Will Thailand commence parliamentary elections on or before 2 February 2014?
77. Will inflation in Japan reach 2 percent at any point before 1 April 2014?
78. Will the U.N. Security Council approve a U.N. peacekeeping operation for the Central African Republic before 1 April 2014?
79. Will negotiations on the TransPacific Partnership (TPP) officially conclude before 1 May 2014?
80. Will Viktor Yanukovich vacate the office of President of Ukraine before 10 May 2014?
81. Will Ukraine officially declare a state of emergency before 10 May 2014?
82. Will there be a lethal confrontation between national military forces from China and Japan before 1 May 2014?
83. Before 1 May 2014, will China confiscate the catch or equipment of any foreign fishing vessels in the South China Sea for failing to obtain prior permission to enter those waters?
84. Before 1 May 2014, will Iran install any new centrifuges?
85. Will there be a significant attack on Israeli territory before 10 May 2014?
86. Will the Israeli-Palestinian peace talks be extended beyond 29 April 2014?
87. Before 1 April 2014, will the government of Venezuela officially announce a reduction in government subsidies for gasoline prices?
88. Before 1 May 2014, will Kenneth Bae leave North Korea?
89. Before 1 May 2014, will China attempt to seize control of Zhongye Island?
90. Will the Bank of Japan (BoJ) officially announce an enhancement of its quantitative and qualitative monetary easing (QQE) policy before 10 May 2014?
91. Will the European Central Bank (ECB) officially announce a plan to charge a negative interest rate on funds parked overnight at the ECB before 31 March 2014?
92. Will Pakistan and the TTP reach a peace agreement before 10 May 2014?
93. Before 1 March 2014, will Russia purchase any additional Ukrainian government bonds?
94. Will family reunions between South and North Korea begin on or before 25 February 2014?
95. Before 1 May 2014, will North Korea conduct a new multistage rocket or missile launch?
96. Will Syria's mustard agent and key binary chemical weapon components be destroyed on or before the 31 March 2014 deadline established by the Executive Council of the Organization for the Prohibition of Chemical Weapons (OPCW)?
97. Will the U.N. Human Rights Council (UNHRC) adopt a resolution directly concerning Sri Lanka during its 25th regular session in March 2014?
98. Will Argentina, Brazil, India, Indonesia, Turkey, and/or South Africa impose currency or capital controls before 1 May 2014?
99. Will the European Union and/or the U.S. impose new sanctions on Viktor Yanukovich and/or members of his government before 10 May 2014?
100. Will Recep Tayyip Erdogan vacate the office of Prime Minister of Turkey before 10 May 2014?

101. Will there be a significant lethal confrontation between armed forces from Russia and Ukraine in Crimea before 1 April 2014?
102. Will Russian armed forces invade or enter Kharkiv and/or Donetsk before 1 May 2014?
103. Will Bahrain, Egypt, Saudi Arabia, or the United Arab Emirates return their ambassadors to Qatar before 10 May 2014?
104. Before 31 December 2014, will China seize control of Second Thomas Shoal?
105. Before 1 May 2014, will the government of Myanmar sign a nationwide ceasefire agreement with the Nationwide Ceasefire Coordination Team (NCCT)?
106. Will Parti Quebecois hold a majority of seats in the Quebec legislature after the 2014 provincial election?
107. Will a referendum on Quebec's affiliation with Canada be held before 31 December 2014?
108. Will China's official annual GDP growth rate be less than 7.5 percent in Q1 2014?
109. Before 10 May 2014, will Russia agree to conduct a joint naval exercise with Iran?
110. Between 2 April 2014 and 10 May 2014, will Russia officially annex any additional Ukrainian territory?
111. Will Iran and the P5+1 countries officially announce an agreement regarding the Arak reactor before 10 May 2014?
112. Will Nouri al-Maliki's State of Law bloc win more seats than any other entity in the 2014 parliamentary elections in Iraq?
113. Will Iran and Russia officially sign an agreement regarding the exchange of oil for goods and services before 10 May 2014?

https://doi.org/10.1017/S1930297500003375 Published online by Cambridge University Press