Public Opinion Quarterly, Vol. 73, No. 5 2009, pp. 895–916

FORECASTING ELECTIONS
COMPARING PREDICTION MARKETS, POLLS, AND THEIR BIASES

DAVID ROTHSCHILD

Abstract Using the 2008 elections, I explore the accuracy and informational content of forecasts derived from two different types of data: polls and prediction markets. Both types of data suffer from inherent biases, and this is the first analysis to compare the accuracy of these forecasts adjusting for these biases. Moreover, the analysis expands on previous research by evaluating state-level forecasts in Presidential and Senatorial races, rather than just the national popular vote. Utilizing several different estimation strategies, I demonstrate that early in the cycle and in not-certain races debiased prediction market-based forecasts provide more accurate probabilities of victory and more information than debiased poll-based forecasts. These results are significant because accurately documenting the underlying probabilities, at any given day before the election, is critical for enabling academics to determine the impact of shocks to the campaign, for the public to invest wisely and for practitioners to spend efficiently.

Starting in the 2008 Presidential campaign, Nate Silver's FiveThirtyEight.com revolutionized election forecasting for the general public. Until his website was launched in March of 2008, those interested in predicting election outcomes typically reviewed national polling results that asked a representative cross-section of voters who they would vote for if the election were held that day. Yet, these raw poll numbers are volatile, subject to random sampling error on either side of the true underlying value. For example, on the eve of the

DAVID ROTHSCHILD is a Ph.D. candidate in the Department of Business and Public Policy (Applied Economics), The Wharton School, University of Pennsylvania, Steinberg-Dietrich Hall, Suite 1400 Philadelphia, PA 19104, USA. I thank my advisor Justin Wolfers for his helpful guidance. Special thanks to my three anonymous reviewers and the editor, Sunshine Hillygus, for their thoughtful and constructive criticism. Also, for all those who provided me feedback: Greg Bilton, Bob Erikson, Wendy Heltzer, Sarah Iosifescu, Adam Isen, Janet Pack, Andrew Paciorek, Betty Rothschild, Dick Rothschild, James Rothschild Jr., Erik Snowberg, Richard Thaler, Christopher Wlezien, and Eric Zitzewitz, thank you for your insightful comments. Address correspondence to David Rothschild; e-mail: RothscDM@Wharton.UPenn.edu.

doi:10.1093/poq/nfp082

2008 Presidential election, national polls showed Obama's lead over McCain ranging anywhere from 2 to 11 percentage points. Starting in the 2000 election cycle, poll aggregation organizations made an improvement by publishing less volatile averages of raw polls; the leading poll aggregators, Pollster.com and RealClearPolitics.com, both had final averages showing Obama winning by 7.9 percentage points over McCain (the final margin was 7.4 percentage points).[1] Although an improvement over raw poll numbers, these estimates still succumb to two well-known poll-based biases, especially earlier in the cycle: polls demonstrate larger margins than the election results and they have an anti-incumbency bias (i.e., early leads in polls fade toward Election Day and incumbent party candidates have higher vote shares on Election Day than their poll values in the late summer into the early fall).[2] Further, they do not provide a probability of victory. In contrast, FiveThirtyEight aggregates raw poll numbers, debiases them toward expected vote share, and then produces a probability of victory. After FiveThirtyEight's strong showing in the Presidential primaries, the discussions of political junkies around the country quickly transformed from focusing on the latest polls to the probability of victory.

Less heralded by the public, prediction markets have been providing probabilities of victory since well before FiveThirtyEight. The Iowa Electronic Market launched the modern era of prediction markets in 1988, introducing a winner-takes-all market in 1992. This type of market trades binary options which pay, for example, $10 if the chosen candidate wins and $0 otherwise. Thus, if there are no transaction costs, an investor who pays $6 for a "Democrat to Win" stock and holds the stock through Election Day, earns $4 if the Democrat wins and loses $6 if the Democrat loses. In that scenario, the investor should be willing to pay up to the price that equals her estimated probability of the Democrat winning the election. The market price is the value where, if a marginal investor were willing to buy above it, investors would sell the stock and drive the price back down to that market price (and vice-versa if an investor were willing to sell below it); thus, the price is an aggregation of the subjective probability beliefs of all investors. Scholars have found that prediction markets are a reliable forecaster in the last few cycles.[3] This is true even though raw prediction market prices experience what is known as the favorite-longshot bias, which drives prices away from probabilities at the tails (i.e., a mean probability of 95 may translate into a price of 85).[4]

[1] Unless noted, all margins are calibrated as the two-party total margin (i.e., if Candidate A has 52 percent support and Candidate B has 44 percent support, Candidate A's margin is $\frac{52-44}{96} =$ 8.3 percentage points).

[2] Campbell (2000), described in next section.

[3] Like much of the previous research, the analysis in this paper relies on Intrade. Intrade is used exclusively because, unlike its competitors, it has markets for all of the Presidential and Senatorial races. In the 2004 Presidential race, it had a greater than 50 percent chance of victory for the winning candidate, on the eve of the election, in all 51 sovereignties.

[4] Leigh et al. (2007), described in next section.

Motivating the analysis in this article is a basic question: are polls or prediction markets more accurate in forecasting elections when the biases of both approaches are corrected? The answer is crucial for researchers studying electoral politics because accurate forecasts allow them to connect shocks to the campaign with changes in the underlying probability of victory, as well as for those studying forecast techniques in a wide range of fields besides politics. It is meaningful to the media that wishes to bring the public the best forecasts, especially as the public decides when and where to invest its time, attention, and money. Finally, better forecasts can help practitioners make more efficient choices when they spend money in the multi-billion dollar industry of political campaigns.[5]

My analysis finds that in the 2008 election cycle FiveThirtyEight's debiased poll-based forecasts were, on average, slightly more accurate than Intrade's raw prediction market-based prices.[6] But when prediction markets are properly debiased, they are more accurate and contain more information than debiased polls; this advantage is most significant for forecasts made early in the cycle and in not-certain races (i.e., the races typically of most interest).

Background

A succession of papers in economics, law, political science, and the popular press have concluded that raw prediction market prices are more accurate predictors of election outcomes than raw polls. The earliest empirical papers originate from studies of the Iowa Electronic Market, with Berg et al. (2001) demonstrating that prediction markets outperform polls in predicting vote share.

The literature is conclusive that polls suffer from biases. Campbell (2000) illustrates the polls' two biases with a chart of the final incumbent party candidate vote share from 1952–1996 on the y-axis and the early-September national polls for the incumbent party candidate on the x-axis. The slope of the regression line is 0.55, demonstrating that leads evaporate by nearly a half. The anti-incumbency bias is demonstrated with the regression line crossing through 50 percent vote share when the poll value is 47 percent (i.e., an incumbent with a poll value of 47 percent receives 50 percent of the vote in expectation).

[5] $1.76 billion was spent on the 2008 Presidential election with an additional $0.94 billion and $0.43 billion on House and Senate elations respectively. The data is from http://www.opensecrets.org/overview/index.php.

[6] Thus, the public made the correct choice, between the readily available forecast options, by predominantly utilizing FiveThirtyEight. Starting in late September of 2008, FiveThirtyEight's page views jumped from about 2× that of Intrade's to over 7.5×; on Election Day FiveThirtyEight had an astonishing five million page views. Please see figure A1 for chart of page views. Data is from http://www.alexa.com/siteinfo/fivethirtyeight.com and http://www.nytimes.com/2008/11/10/business/media/10silver.html?scp=5&sq=Stephanie%20Clifford&st=cse.

At the same time, a separate literature has shown that prediction market prices also suffer from inherent biases. In a theoretical paper, Wolfers and Zitzewitz (2004, p. 108) assert that "In a truly efficient prediction market, the market price will be the best predictor of the event, and no combination of available poll or other information can be used to improve the market-generated forecast." Manski (2005) highlights Wolfers and Zitzewitz's "efficiency" caveat and demonstrates theoretically that issues regarding the risk profile of the traders distort the translation of investors' mean probability beliefs into prices. Wolfers and Zitzewitz (2007) show that in addition to nonrisk neutral investing, the favorite-longshot bias inherent to prediction markets is caused by transaction costs and liquidity concerns. To illustrate this bias, assume that an investor believes the Democrat has a 95 percent chance of winning sixty days before the election. Because of the opportunity costs of the bet being held for 60+ days (there is limited liquidity in many markets) and transaction costs of $0.015 per $1.00, the investor will actually bid up to only about $0.85 per $1.00, rather than $0.95 per $1.00. Further, if there are two bets that are equal in expectation, the investor gains more utility from betting on a longshot.[7] The bias is documented empirically in Leigh et al. (2007).[8] (I am not testing the efficient market principle in this paper, but I accept that arbitrage is not capable of overcoming the inherent bias in this particular market.)

In a recent paper in this journal, Erikson and Wlezien (2008) advance the debate between polls and prediction markets when they argue that while raw prediction market prices may provide more accurate forecasts than raw polls, adjusting the polls for known biases reverses this result. Thus, they argue that "market prices contain little information of value for forecasting beyond the information already available in the polls" (2008, p. 24). The problem is that Erikson and Wlezien do not advance the literature far enough. Their paper is the first empirical comparison that includes debiased polls and it is the first to focus on probability of victory, rather than just expected vote share. But the authors treat the well-documented favorite-longshot bias in prediction markets as a weakness of the markets rather than a systematic bias that can easily be corrected. In their conclusion they note the persistent problem of "The winner-takes-all market . . . overvaluing longshot candidates' chances of victory" (p. 24).

This analysis extends the literature in three main ways. First, it debiases both prediction market and polling forecasts when comparing their accuracy. Second, it updates Erikson and Wlezien's approach so that it can be applied to state-level races and consequently evaluates a much larger sample of elections. Finally, it utilizes a more sophisticated transformation of the raw polls that improves upon Erikson and Wlezien's method, while maintaining its general

[7] Neither Manski nor this paper conclude whether investors are risk loving or beset by misconceptions or Prospect Theory, but it is accepted in the literature that they are not risk neutral.

[8] A preliminary version of Leigh et al. (2007) was presented at the 2007 UC Riverside conference on Prediction Markets.

structure. This new method debiases and then transforms the poll aggregation values, while Erikson and Wlezien debiases and then transforms the raw poll numbers. Using Erikson and Wlezien's method, the probability in some races swing an implausible 30–40 percentage points around a trend on a daily basis, making it of little use for real-time predictions.[9] Thus, this approach is both more realistic and easier to compare with prediction markets. In addition, the analysis also tracks the forecasts of FiveThirtyEight, the best-known poll forecaster. Although the method used by FiveThirtyEight is somewhat opaque, it offers an interesting comparison to the other forecasts. Since FiveThirtyEight reports its probabilities in real time, there is no concern of inadvertent look-ahead bias that could afflict forecasts created ex-post.

Data

The analysis examines seventy-four races over the last 130 days of the 2008 campaign: fifty Presidential Electoral College races and twenty-four contested Senatorial races. This is in contrast to the four national Presidential races (1992–2004) reviewed by Erikson and Wlezien (2008). None of the seventy-four elections are completely independent; there are national as well as regional trends that affect several to all of the polls at one time. Yet, on any given day, the seventy-four different forecasts represent seventy-four different decisions about how to weigh the interdependent data and thus provide more information about the accuracy and information inherent to the different types of forecasts than four races in different cycles.

The first step in creating a poll-based forecast is to create a snapshot, which is the estimated two-party vote share of the two candidates if the election were held that day. The Erikson and Wlezien method, labeled as Poll_EW, uses the latest poll at any given day before the election as its snapshot; for this method, I pool the polls if there are more than one and use the most recent if there are none that day. The new method, noted as Poll_Debiased, creates a linear regression of all polls up to that day, and the snapshot is the trend of that regression.[10] FiveThirtyEight weighs all polls by pollster, sample size and recentness and then adjusts that average for national trends. The snapshot is completed by adding a regression of expected vote share on demographic and historical political data, which is weighted heavily in the snapshot only when there is insufficient polling data available.

The second step in creating a poll-based forecast is to create a projection, which is the estimated vote share of the two candidates on Election Day. To

[9] Please see figure A2 for a chart of Erikson and Wlezien (2008)'s method applied to the 2008 race. I have also adapted it for state races in figure 1.

[10] Poll aggregators create a snapshot using a combination of averages, linear trends, and/or loess trends. I use just the linear trend, because it the simplest and most transparent method to create a consistent poll average on any given day, especially in races with limited number of polls.

create the projection of both Poll_EW and Poll_Debiased, I regress the final vote share on the poll for each day before the election in previous election years: $V_{yr} = \alpha + \beta P_{yr} + e_{yr}$, where y is a given year and r is a given race. All transformations are optimized with out of sample data: elections from 2000 and 2004 for the Presidential races and 2004 and 2006 for the Senatorial races.[11] I recover a unique alpha and beta for each day before the election (T), and the daily projections for 2008 are created using those parameters: $\hat{V}_{2008,T} = \alpha_T + \beta_T P_{2008,T}$; the alpha corrects for the anti-incumbency bias and the beta corrects for reversion to the mean. For Presidential races, FiveThirtyEight projects the snapshot using historical trends of national poll movement and their correlation to the individual states. Undecided voters are allocated 50/50 to the major candidates after the third party probable vote share is taken out. For Senatorial races, FiveThirtyEight uses the snapshot as the projection.

The third step in creating a poll-based forecast is to create a probability of victory, which is the probability that the estimated vote share is greater than 50 percent. Poll_EW and Poll_Debiased model the vote share on Election Day as a normal distribution around the projection. For the same projection, the more accurate the estimation of the projection is, the tighter the distribution, and the greater the percentage of probable outcomes where the favored candidate has the higher amount of votes. Mimicking Erikson and Wlezien (2008), Poll_EW assumes that the accuracy of the projection decreases with the accuracy of the estimated vote totals and the distance of 2008's poll from the average poll at this point in the election cycle. Thus, the probability of victory originating from any given day before the election can be estimated as follows: $Pr = \Phi(\hat{V}_{2008,T}/RMSE_T + V(\beta_T)(P_{2008,T} - Pr))$. For Poll_Debiased I use maximum likelihood to determine the optimal sigma (σ_T) for each day: $Pr = \Phi(\hat{V}_{2008,T}/\sigma_T)$.[12] FiveThirtyEight simulates the data with a Monte Carlo analysis 10,000 times. The simulation accounts for: sampling error, state-specific and national movement. The probability of victory is the percentage of simulations that the candidate gets over 50 percent of the vote.[13]

The prediction market data, from Intrade, needs to be translated from prices into probabilities. First, I take the average of the bid and ask for the stock that pays out if the Democrat wins on Election Day. If the bid-ask spread is greater than five points, I take the last sale price.[14] If there are no active offers and no sales in the last two weeks of the race, I drop the race; this includes

[11] The data is collected from: PollingReport.com, Pollster.com, and RealClearPolitics.com. Using the method from Erikson and Wlezien (2002) I fill in missing data, for historical data only, with the linear interpolation from the poll before and after any missing day.

[12] For all of Poll_Debiased's parameters I use ± 7 days of data to gain consistency, relative to the daily random variation in the Erikson and Wlezien model.

[13] The formation of FiveThirtyEight's probability is explained in more detail at: http://www.fivethirtyeight.com/2008/03/frequently-asked-questions-last-revised.html. It is possible that it updated its method during the cycle.

[14] Procedure is adapted from Snowberg et al. (2007).

one Presidential race (DC) and eleven Senatorial races (AL, AR, DE, IA, IL, MI, MT, RI, TN, WY.I, and WY.II).[15] The data recovered from these first two steps I refer to as Raw Intrade. To correct the favorite-longshot bias I use the transformation suggested by Leigh et al. (2007): $Pr = \Phi(1.64^{*}\Phi^{-1}(price))$.[16] I refer to this forecast as Debiased Intrade.

The five forecasts are compared for their value during the last 130 days of the cycle (i.e., June 27 through Election Day 2008). The methods for Poll_EW and Poll_Debiased provide one probability per day; I date a poll as being released the day after its final day in the field. FiveThirtyEight updated its Presidential probabilities regularly since March 2008 and published ninteen rounds of forecasts for Senate races. I use all 14 different rounds of Presidential forecasts that I have been able to obtain and all 19 Senatorial forecasts.[17] When FiveThirtyEight is compared directly with any of the other forecasts, I use the other forecasts' closest previous forecast. I use Intrade numbers from noon on each day.[18]

Figure 1 shows the progression of Poll_EW and Poll_Debiased's probabilities of victory for the incumbent party candidate, Republican John McCain, for the national popular vote over the course of the campaign; the left side of figure 1 demonstrates why I drop Poll_EW moving forward. The shifts in the underlying probability of victory cannot justify the volatility in Poll_EW. Thus, Poll_Debiased is specifically created to be a more realistic and less volatile version of Poll_EW; the chart illustrates how Poll_Debiased exists near the mean of Poll_EW's trend. Moreover, as a practical implication, Poll_EW is so volatile that it is hard to grasp anything else on a graph that includes it.

The right side of figure 1 is the same as the left side of figure 1, but excludes Poll_EW and adds Raw and Debiased Intrade's forecasts as well as annotations of the major events from the election cycle; this side of the figure illustrates the value of determining the underlying probability of victory. The chart demonstrates that while there is strong correlation between the polling and market-based forecasts (the Intrades are tied together by construction), there is still considerable variation at points during the cycle. Both Poll_Debiased and Intrade have McCain moving upward after the Republican National

[15] An example of "no active offers" would be an investor willing to buy at 92, but no investor willing to sell. Eleven of the twelve dropped races have negligible volume for the entire cycle, with WY.I having twenty outstanding shares, but no additional volume after June and no bid or ask down the stretch.

[16] This transformation was suggested (and estimated) prior to my sample, using data from Presidential predication markets from 1880 to 2004. The process for determining 1.64 is the same as my eq. (1) later in the paper. The authors take the inverse normal of all of the prices they collected and then solve for the coefficient of the data in a probit.

[17] The fourteen Presidential forecasts are what Peter McCluskey of the BayesianInvestor.com and I randomly saved, which are disproportionately later in the cycle. FiveThirtyEight has not responded to my request for further historical data and it is not available at Archive.org.

[18] Unfortunately, there are ten random days where I do not have Intrade data; those days are dropped from the direct comparisons with Poll_EW and Poll_Debiased.

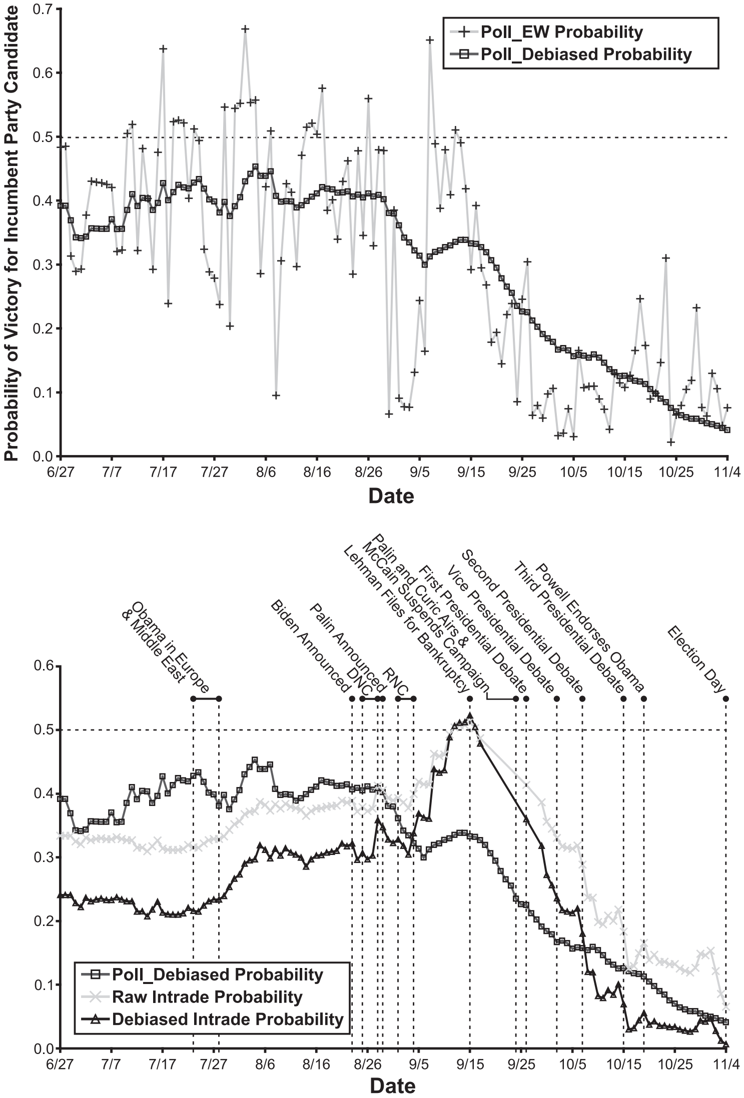

Figure 1. Probability of Victory in the National Popular Vote for the Incumbent Party Candidate in the 2008 Presidential Election.

NOTE.—The incumbent party candidate, Republican John McCain, lost by a margin of 7.4 percentage points in the votes cast for the two major party candidates.

Convention and the announcement of Sarah Palin as his running mate, but only Intrade has him crossing the 50 percent threshold (i.e., predicting he wins the election). Yet, even if there was a consensus on the underlying national values, it is impossible to determine causality of events on outcomes using national data calibrated daily; there are too few races, just one every four years, and too many overlapping events. Thus, extending forecast research to state-level races is essential to gathering the data necessary determine some causality or, at minimum, a fuller description of correlations between events and electoral outcomes. Of course another important reason for focusing on state-level races is that the national popular vote does not determine the winner of the U.S. Presidential election since the election outcome hinges on the results in fifty-one individual sovereignties, through the Electoral College.[19]

Methods/Results

Previous research offers a variety of techniques for evaluating the accuracy of a forecast. The most simplistic approach is to determine how often each forecast correctly predicts the winning candidate with probability of victory of at least 50 percent–a basic threshold accuracy measure. This metric, however, offers little leverage in comparing the forecasts. On the eve of Election Day, all of the forecasts have >50 percent probability of victory for all of the Senatorial winners and they are all >50 percent for forty-nine or fifty of the fifty-one Presidential winners. Even looking at mid-September projections, the forecasts are nearly indistinguishable, with between seventy-six and seventy-eight of eighty-six favoring the winning candidate. By construction, the Intrades are the same using the 50 percent threshold measure of accuracy and there are just 473 observations, out of 8,361 total observations, where Poll_Debiased and Intrade differ; Intrade is correct in two-thirds of the observations. The distinction between FiveThirtyEight and Intrade is even smaller and less significant. There are twenty-five observations out of 1,156 where the two forecasts differ based on the 50 percent threshold measure of accuracy; these observations are clustered within a few races and have one forecast just over 50 percent and the other just below 50 percent.[20] Thus, in order to evaluate the forecasts, I need to examine the distribution of their probabilities, not just whether the favored candidate won.

The charts in figure 2 show the percentage of the forecasts' probabilities of victory for the winning candidate, on a given day before the election, which reach the following thresholds: >90 percent, >50 percent, and >25 percent.

[19] There is also evidence that the national popular vote prediction markets may suffer from manipulation by people motivated to gain publicity for their chosen candidate, but this evidence does not extend to the state-level markets.

[20] FiveThirtyEight's lowest probability of victory for the winning candidate in these disagreeing observations is 34 percent and Debiased Intrade's is 30 percent, while their highest is 74 percent and 80 percent, respectively.

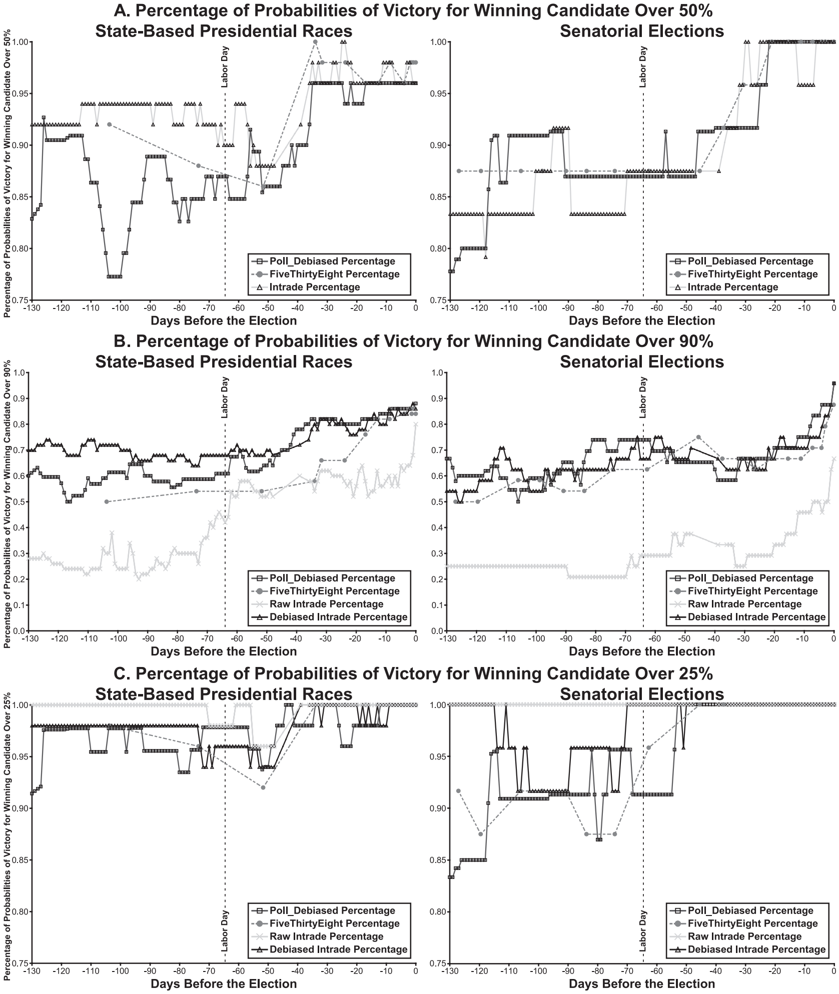

Figure 2. Distribution of Probabilities of Victory for Winning Candidates. NOTE.—This figure shows the percentage of the forecasts' probabilities of victory for the winning candidate, on a given day before the election, which reach the thresholds noted on each chart. On the left side of the figure the percentage is from the fifty Presidential Electoral College races in the sample; on the right side, it is the from twenty-four Senatorial elections.

These charts help illustrate the sources of identification that underlie the subsequent accuracy metric that will be used to evaluate the different forecasts.

In the Presidential races prior to Labor Day, Intrade is stronger at keeping predictions for winning candidates >50 percent relative to Poll_Debiased and FiveThirtyEight, while the forecasts are very competitive after Labor Day

(2A). There is little distinction between the forecasts in the Senatorial races, except for a few observations, well before Labor Day, where FiveThirtyEight and Intrade have a persistent difference in their forecast for the special Mississippi Senate race. Between the poll-based forecasts, FiveThirtyEight is stronger than Poll_Debiased in forecasts above 50 percent for the Presidential races.

In probabilities above 90 percent FiveThirtyEight is a cautious predictor, moving toward these more certain probabilities late in the cycle; Poll_Debiased and Debiased Intrade are similar to each other, with Debiased Intrade showing slightly more confident probabilities early in the Presidential cycle (2B). Since too few of the candidates with 80–90 percent probability of victory lose on Election Day, especially for FiveThirtyEight, the more observations a forecast has >90 percent, the more accurate its depiction of the true underlying probabilities. FiveThirtyEight's overly cautious predicting is most extreme in the Presidential races and persistent, but smaller, in the Senatorial races. All of the forecasts demonstrate less confidence in their Senatorial versus Presidential predictions. They state probabilities >90 percent less often and increase the percentage of forecasts in the top thresholds later in the cycle. The relative uncertainty in the Senatorial races is increased because the Senate accounts for eleven of the twelve highly certain races dropped due to lack of Intrade data. Due to its favorite-longshot bias, Raw Intrade has very few of these extremely confident forecasts.

Debiased Intrade and FiveThirtyEight give the winning candidate little chance to win earlier in the cycle, and less randomly, than Poll_Debiased. As the cycle progresses and the amount of potential shocks to the races decrease, fewer and fewer observations should fail to reach the >25 percent threshold (2C). FiveThirtyEight has some very wrong predictions in the first half of the cycle, especially in Senatorial races, but refrains from predicting the winner with <25 percent by mid-September (at that point any missed observations are approaching 50 percent). Poll_Debiased, not benefiting from FiveThirtyEight's regressions, has very wrong predictions early in the Senatorial and Presidential cycle and not benefiting from weighing the polls, continues to produce randomly wrong forecasts much later in the cycle. Yet, toward the middle of the cycle, FiveThirtyEight and Debiased Intrade both overcompensated for the postconvention Republican bounce, where Poll_Debiased benefits from relying only on present polls and not estimating future movement. With its lack of confidence, Raw Intrade avoids both extremes and has few extremely wrong predictions, none in the Senatorial races.

The most interesting forecasts are in races in which the outcome is not-certain and for subsequent analysis I define observations with probabilities >90 percent as observations where the forecasts are in the "certain" range.[21] In all of FiveThirtyEight and Debiased Intrade's forecasts, just 0.27 percent of

[21] Any finding in this paper relating to the 90 percent line is robust to nearby probabilities.

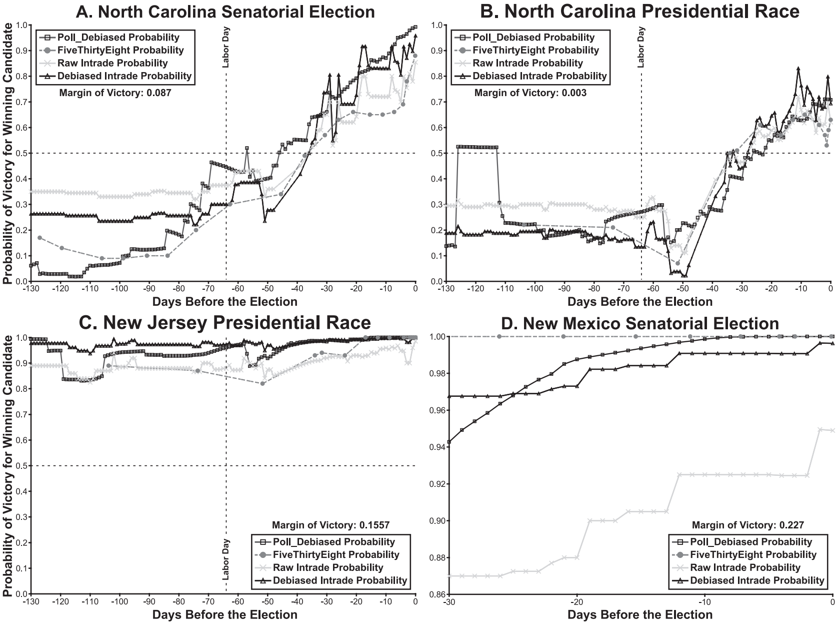

Figure 3. Probability of Victory for the Winning Candidate in Select Races.

candidates lost when they were predicted to win with >90 percent probability.[22] In many respects these certain races are not as important to the observers of elections because they are perceived as decided races and, thus, distinctions between the forecasts are more arbitrary, as there are fewer polls and fewer participants in the prediction markets. Most races eventually become certain by the end of the campaign. For example, on Election Eve FiveThirtyEight had just seven Presidential contests that were not >90 percent probability of victory for one candidate: FL, IN, MO, MT, NC, ND, and OH. On the earliest day in my sample FiveThirtyEight had twenty-four Presidential races that were not >90 percent.[23]

Figure 3 shows the progression of the forecasts for the probability of victory for the winning candidate in four individual contests which were chosen to demonstrate persistent trends that contribute to the aggregated accuracy of the forecasts. First, the two North Carolina races highlight the fact that using a

[22] FiveThirtyEight and Debiased Intrade have nineteen observations, out of over 9,500 total observations, where they give >90 percent probability of victory to the eventual losing candidate. These observations are in the NC Presidential and Senatorial races and the IN Presidential race. The only observation where both FiveThirtyEight and Debiased Intrade gave the eventual winner a less than 10 percent probability is NC's Presidential race two days before Lehman collapsed.

[23] On Election Eve, FiveThirtyEight had just three Senate race where there was not >90 percent probability of victory for one candidate: GA, MN, and NC. On the earliest day in my sample FiveThirtyEight had twelve Senate races that were not >90 percent.

50 percent threshold as an accuracy benchmark would miss critical information in the forecast because the forecasts cross the 50 percent line nearly in tandem. Second, the charts show that Poll_Debiased and Debiased Intrade are generally quicker to cross over into the >90 percent range than FiveThirtyEight. Further, they both show more confidence than FiveThirtyEight in the expected and eventual winner, on average, in races like the North Carolina Presidential race, which are still uncertain on Election Day. Third, while Debiased Intrade is a little less likely to get severely wrong predictions than FiveThirtyEight, both of them bottom out badly about a week after the Republican National Convention, as shown in the two North Carolina races. Fourth, Raw Intrade is the most conservative; it avoids large mistakes, but also shows less confidence in certain races. Fifth, I generally find that Debiased Intrade slightly trails the poll-based forecasts in the most certain races, such as the New Mexico Senate race.

The races in figure 3 also illustrate that the larger or most persistent differences between the forecasts do not necessarily come from the more competitive races or the even the biggest moments a race. First, margin of victory does not necessarily determine if the race was certain or uncertain during the course of the cycle. Democrat Kay Hagen easily won the North Carolina Senate race with 53 percent of the vote over Republican Elizabeth Dole's 44 percent of the vote (a 0.087 margin). Yet, despite this comfortable margin of victory, the race was far from certain to forecasters for most of the cycle. By comparison (not shown) forecasters (and most observers) forecast McCain to carry his home state of Arizona with >90 percent probability throughout most of the cycle, even though he won with a margin of 0.086, a slightly smaller margin than the Democrat's victory in North Carolina's Senate race. Second, the uncertain races do not necessarily provide much identification in terms of absolute difference. North Carolina was a competitive state through the entire Presidential campaign, with Barack Obama squeaking out a win with a tiny 0.003 margin of victory. But even though North Carolina was an uncertain for the entire campaign, the different forecasts never stray too far apart. In contrast, New Jersey, long a Democratic stronghold in Presidential politics, was easily won by Barack Obama 57 percent to McCain's 42 percent (a 0.157 margin). Despite this large margin of victory, the absolute difference between the forecasts is actually more extreme at points in New Jersey's Presidential race than in the competitive North Carolina race.[24]

Since distinctions within the certain range (i.e., probabilities >90 percent) are arbitrary, not as interesting as distinctions in the uncertain range and common in the data, it is important that the scoring rule for accuracy puts more weight on the failure to correctly place forecasts in the 90–100 percent range versus the 80–90 percent range, than correctly placing forecasts in the

[24] Please see the supplementary data online (figure A4) for charts with the progression of the Florida and Ohio Presidential races, both for the general interest in those races and as a useful comparison to show the heightened movement, and hence benefit in determining correlation and causality, in the state-level markets relative to the national market.

95–100 percent range versus 90–95 percent range; the mean difference of the square error (MSE), the main scoring rule used to quantify the accuracy of the forecasts, does this.[25] The square error is $(1 - Prob(Victory))^2$, where $Prob(Victory)$ is for the winning candidate. A strategic forecaster maximizes his score, in expectation, by forecasting his true belief. To illustrate what the attributes of the scoring rule mean in this article, in the lower half of the certain range only 1 percent of Debiased Intrade's 1,013 forecasts between 90–95 percent lose, and 2 percent of FiveThirtyEight's; thus both forecasts, Debiased Intrade more than FiveThirtyEight, should be more confident with these observations (i.e., move the probability of victory for the chosen candidate closer to 98 or 99 percent). Yet, more heavily weighted by MSE is that 11 percent of Debiased Intrade's probabilities between 80–90 percent lose, while only 7 percent of FiveThirtyEight's probabilities lose in that range (i.e., as illustrated in figures 2 and 3, FiveThirtyEight is leaving many observations in the 80–90 percent range that should be in the 90–100 percent range).

For presidential races, as shown in figure 4, Debiased Intrade has a statistically significant smaller mean square error relative to the poll-based forecasts until mid-September and then continues to have smaller errors until the end of the cycle.[26] The figure charts the mean difference in square error for the two poll-based forecasts relative to Debiased Intrade in the Presidential races; anything above zero indicates a more accurate mean forecast for Debiased Intrade. In the beginning of the cycle, Debiased Intrade is slightly besting FiveThirtyEight in races of all degrees of certainty, while it is beating Poll_Debiased on the most uncertain races and doing similarly for the vast majority of races. This translates into a modest advantage over FiveThirtyEight and a commanding lead over Poll_Debiased, as MSE puts more emphasis on uncertain observations. Toward the middle of the cycle, Poll_Debiased is able to pull ahead of Debiased Intrade, because Debiased Intrade and FiveThirtyEight were making a few massive mistakes in this time period. Finally, toward the end of the cycle, Debiased Intrade has a slight advantage over FiveThirtyEight; the main identification at the end is Debiased Intrade having more confidence in the not-certain races and FiveThirtyEight demonstrating more confidence in the most certain races. Poll_Debiased falls far behind the other forecasts, because it is the only one still making massively wrong predictions. I only show the chart with Debiased Intrade, because it is the more accurate of the

[25] Please see the supplementary data online (figures A5 and A6) for charts regarding an alternative approach, the mean difference of the absolute error (MAE), $(1 - Prob(Victory))$. This scoring rule rewards a forecaster equally if he forecasts 75 to 70 percent as it would if he forecasts 95 to 90 percent. Thus, a strategic forecaster can maximize his expected score by stating 100 percent probability for any candidate with >50 percent probability of victory. So while MSE is driven by the distinctions in the important and precisely calibrated observations, MAE, especially later in the cycle, is driven by the differences among less important and less precisely calculated observations.

[26] While I view the standard errors as a lower bound, due to issues involving the independence of the forecasts, I believe that the statistical significance is still a meaningful guide to the degree of differences between the different forecasts.

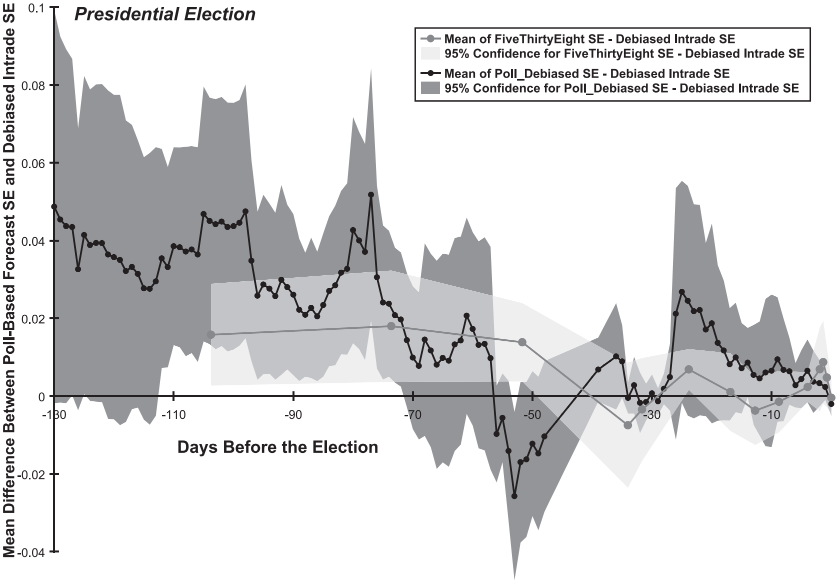

Figure 4. Mean of Poll-Based Forecast's Square Error–Debiased Intrade's Square Error with 95 percent Confidence Interval.
NOTE.—Each point plots the difference in MSE arising from the forecasts issued at that point in time for the fifty Presidential Electoral College races in the sample.

two prediction market forecasts. Please see figure A3 for the chart comparing the Raw Intrade and Debiased poll forecasts. Raw Intrade does not make big mistakes in the most uncertain observations, which helps it have a statistically insignificant, but smaller error than the poll-based forecasts for the first half of the cycle. Lacking confidence in most certain races, it falls behind the poll-based forecasts in the second half of the cycle, consistent with the findings of Erikson and Wlezien (2008). Yet, figure 4 shows that debiasing Intrade creates a forecast with a statistically significant smaller mean square error relative to the poll-based forecasts in Presidential races, especially earlier in the cycle.

In the Senatorial races, Debiased Intrade has a statistically insignificant smaller mean square error relative to FiveThirtyEight and a similar error relative to Poll_Debiased. Figure 5 replicates figure 4 for the Senatorial races, again showing the mean difference in square error for the two poll-based forecasts relative to Debiased Intrade. FiveThirtyEight's persistently (slightly) worse error is due to its several very wrong predictions prior to Labor Day and its consistently fewer predictions over 90 percent toward the end of the cycle.

The figures focus on the accuracy of the forecasts as they are reported, the following tables are designed to consider their informational content. I start

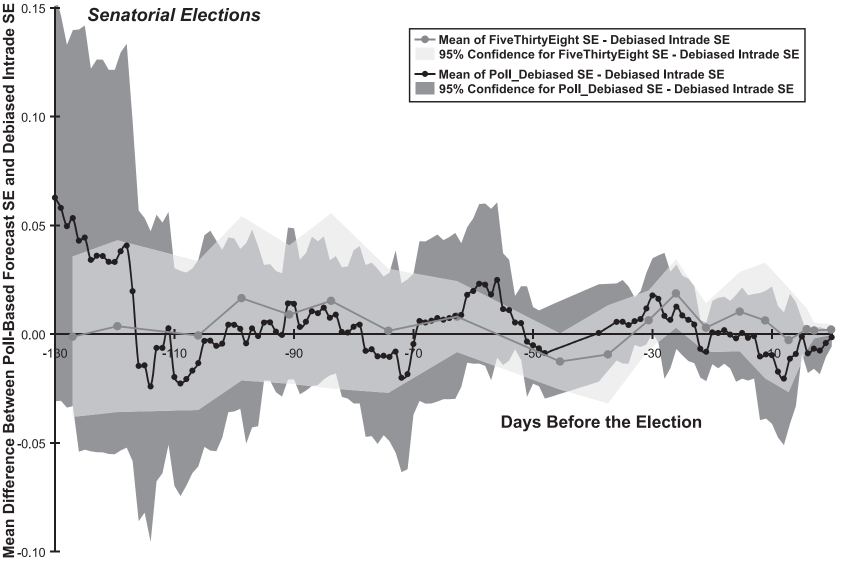

Figure 5. Mean of Poll-Based Forecast's Square Error–Debiased Intrade's Square Error with 95 percent Confidence Interval.
NOTE.—Each point plots the difference in MSE arising from the forecasts issued at that point in time for the twenty-four Senatorial elections in the sample.

with a probit analysis to appraise the confidence level of the forecasts and determine if there are any systematic biases associated with forecasts in the 2008 election. Reported in table 1 are the results from a probit model:

$$I(Win)_r = \Phi(\beta_0 + (1 + \beta_1)^{*}\Phi^{-1}(Forecast)_{r,t}) \tag{1}$$

where $I(Win)_r$ is an indicator variable for whether the noted candidate won race r. If β_1 is positive the forecast would have been more accurate by adding confidence to its probabilities (i.e., the forecast would be more accurate if it made all of its predictions stronger). If β_0 is significant then I cannot rule out that the forecast was systematically favoring one side versus the other or there was a 2008 specific shock in one direction (i.e., the forecast would be more accurate if it systematically moved all of the predictions in one direction). I ran both tables 1 and 2 twice, with the incumbent candidate and then the Republican candidate as the dependent variable; since the results are almost identical I only show the incumbent party candidate as the dependent variable. I do not show any of the results for Raw Intrade, because they are the exactly the same as Debiased Intrade, just multiplied by 1.64.

I have almost all negative and significant constants (rows c, f), which indicates all of the forecasts would have benefited from systematically adding a few points in the direction of the nonincumbent candidate for all of their

Table 1. Coefficients from Probit of Winner on Forecasts, Where the dependent variable Is $I(IncumbentWin)_r$

	Panel I: Poll_Debiased and Intrade	
(a) Poll_Debiased	0.320*	
	(0.143)	
(b) Debiased Intrade		0.659*
		(0.250)
(c) Constant	−0.785*	−0.824*
	(0.247)	(0.364)
Observations	8,361	8,361
	Panel II: FiveThirtyEight and Intrade	
(d) FiveThirtyEight	0.936*	
	(0.278)	
(e) Debiased Intrade		0.662*
		(0.232)
(f) Constant	−0.814*	−0.536
	(0.346)	(0.353)
Observations	1,156	1,156

NOTE.—Standard errors are shown in parentheses and clustered by race: 74 total. *denotes statistical significance at the 5% level.

probabilities. This result is not surprising. All of the transformations are fitted for prior elections, so constants will be significant if 2008 systemically differs from recent years, which it did. Almost every uncertain race broke toward the Democrat or the nonincumbent. While an imperfect arbitrator of uncertainty through the 130 days prior to the election, seven of nine races that were decided by 5 points or less went to the Democrat or nonincumbent.

All of the forecasts are under-confident, but FiveThirtyEight is much more under-confident than Debiased Intrade and Debiased Intrade is much more under-confident than Poll_Debiased. FiveThirtyEight's under-confidence was evident in the earlier analysis in that their predictions left many probabilities short of the >90 percent category well later than other forecasts. For Intrade, it means that the transformation suggested by Leigh et al. was not strong enough for 2008 and that Intrade would have been most accurate if the transformation coefficient was 2.72 versus 1.64.[27] Poll_Debiased's relatively small need for additional confidence is also evident in figures 2 and 3, where it is relatively aggressive in placing observations above 90 percent.

Reported in table 2 are the results from a simple binary test in the spirit of Fair and Shiller (1989 and 1990) to examine whether the forecasts provide

[27] Since Debiased Intrade's $(1 + \beta_1) = 1.660$, Raw Intrade's $(1 + \beta_1) = 1.660*1.64 = 2.72$.

Table 2. Coefficients from Probit of Winner on Forecasts, Where the dependent variable Is $I(IncumbentWin)_r$

	All observations	Before labor day	Not-certain races^
	Panel I: Poll_Debiased and Intrade		
(a) Poll_Debiased	0.451*	0.267	0.357
	(0.214)	(0.228)	(0.219)
(b) Debiased Intrade	1.253*	1.407*	1.153*
	(0.369)	(0.428)	(0.364)
(c) Constant	−0.941*	−0.915*	−0.894*
	(0.350)	(0.393)	(0.310)
Observations	8,361	4,354	3,167
	Panel II: FiveThirtyEight and Intrade		
(d) FiveThirtyEight	0.846	0.514	0.774
	(0.456)	(0.602)	(0.495)
(e) Debiased Intrade	0.981*	1.170	0.960*
	(0.418)	(0.643)	(0.423)
(f) Constant	−0.714*	−0.763	−0.690*
	(0.356)	(0.458)	(0.341)
Observations	1,156	268	380

NOTE.—Standard errors are shown in parentheses and clustered by race: 74 total. *denotes statistical significance at the 5% level.

^I drop races where both forecasts are >90% probability. Roughly two-thirds of the remaining observations occur after Labor Day.

unique information from each other:

$$I(Win)_r = \Phi(\beta_0 + \beta_1{}^*\Phi^{-1}(Intrade)_{s,t} + \beta_2{}^*\Phi^{-1}(PollForecast)_{r,t}) \qquad (2)$$

Whereas the earlier analysis compared the accuracy of the forecasts as they were reported, the results here explore the accuracy of the forecasts, but with an optimal manipulation of the information they provide. The coefficients adjust for any issues in the confidence and bias of the forecasts. This is akin to asking if I were to make a new forecast, optimally combining the forecasts in this study as my raw information, how much of each forecast would be used. There are two things to consider when examining these results: the relative size of coefficients illustrates the weight placed on each forecast and the statistical significance confirms if I can reject that one forecast encompasses all of the useful information in the other. I run this probit for all observations, just observations occurring before Labor Day, and dropping all observations where both forecasts are >90 percent.

I can reject the possibility that Intrade (rows b, e) contains no independently valuable information but I cannot reject, under most circumstances, the possibility that all of FiveThirtyEight (row d) or Poll_Debiased's (row a) information is encompassed by Intrade. In the only category where Intrade is not significant at the 5 percent level, it is significant at the 10 percent level. FiveThirtyEight is never significant and Poll_Debiased is significant only in the all observations category. If I were to make a joint forecast, I would heavily emphasize Debiased Intrade and may be just as accurate without either of the poll-based forecasts.

Conclusion

In 2008, FiveThirtyEight, a debiased poll-based forecast, offered to the general public a more accurate forecast than raw poll numbers or raw prediction market prices. But, the analysis here shows that were Intrade's prices debiased, they would have provided a more accurate forecast and more valuable information than the best poll-based forecasts currently available, especially early in the cycle and in uncertain races. An examination of the structure of these forecasts helps explain this informational advantage.

There are three main components to a forecast: the raw information being aggregated, the transformation of this information into probabilistic forecasts, and any bias that shifts the stated forecasts.

As to information, the raw information used by the poll-based forecasts is public and hence should be in the information set of Intrade investors. Beyond this, prediction markets aggregate dispersed and unpublished information (i.e., a brewing scandal may be known to a few investors before the general public). Also, prediction markets are capable of incorporating new information in real-time, whereas poll-based forecasts take several days for information to saturate (i.e., a publicly-known event is immediately incorporated into the stock price, but it will take several days before it is fully incorporated into the polls). Further, prediction market stocks are based on the value of the candidates on Election Day; thus, investors are incorporating their information on how it will affect the race on Election Day, while poll-based forecasts are only able to debias the information based off of previous cycles (i.e., investors can discount a bump in the polls generated by the visit of a popular leader, but poll-based forecasts can only discount the bump if it happened regularly, at the same time, in previous cycles). FiveThirtyEight supplements its forecasts with its historical regressions when there are few polls, which mitigates the true disadvantage of forecasting with only polls.

As to transformations, the poll-based forecasts have sophisticated methods for transforming information into probabilities. Investors in Intrade vary in their methods of converting information into subjective probabilities and then those probabilities are aggregated by the certainty of the investors. Only a small percentage of the investors will be as sophisticated as the poll-based forecasters and there is no guarantee that they will be the most certain of the investors. As to reporting biases, it is now possible to correct for biases in reporting or look past the biases for the informational content of the forecast.

Since the informational advantage of Intrade's forecasts is not derived from more sophisticated transformations or less biased reporting, it must originate from higher informational content in its raw data. The results of this analysis, and increased knowledge about the structure of these forecasts, can be utilized to make stronger market-based forecasts as well as stronger poll-based forecasts, both inside and outside of politics.

Appendix

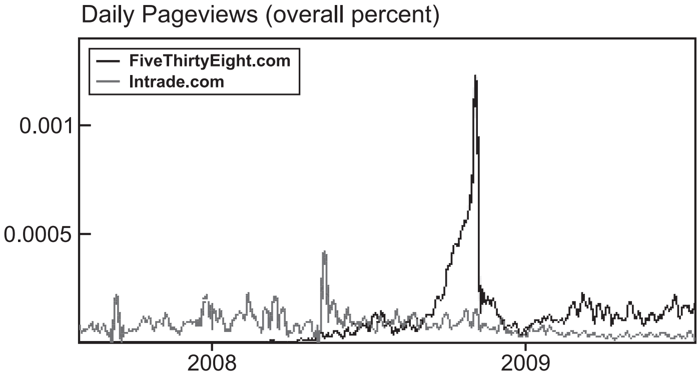

Figure A1. Comparison of Pageviews for FiveThirtyEight.com and Intrade.com During 2008 Election Cycle.
NOTE.—FiveThirtyEight.com launched in March of 2008. Data is from Alexa.com

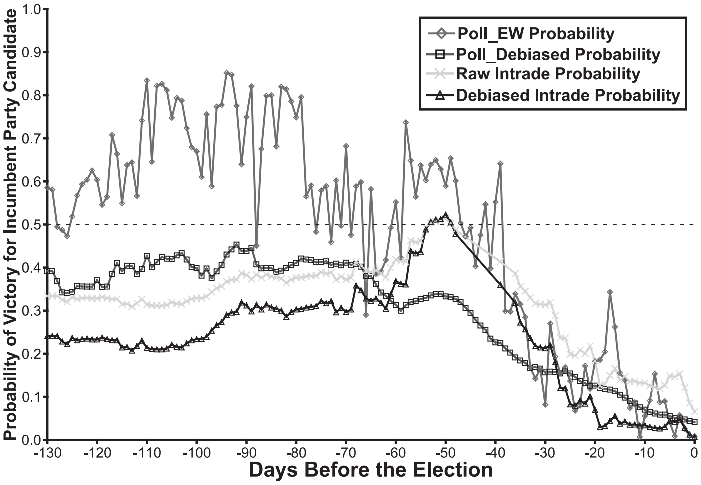

Figure A2. Probability of Victory in the National Popular Vote for the Incumbent Party Candidate in the 2008 Presidential Election
NOTE.—This figure differs from Figure 1 because instead of using data from the 2000 and 2004 state by state Presidential races to create the transformation parameters for Poll_EW, it uses the national data from 1952–2004. The parameters closely resemble those printed in Erikson and Wlezien (2008), which were created with the national data from 1952–2000.

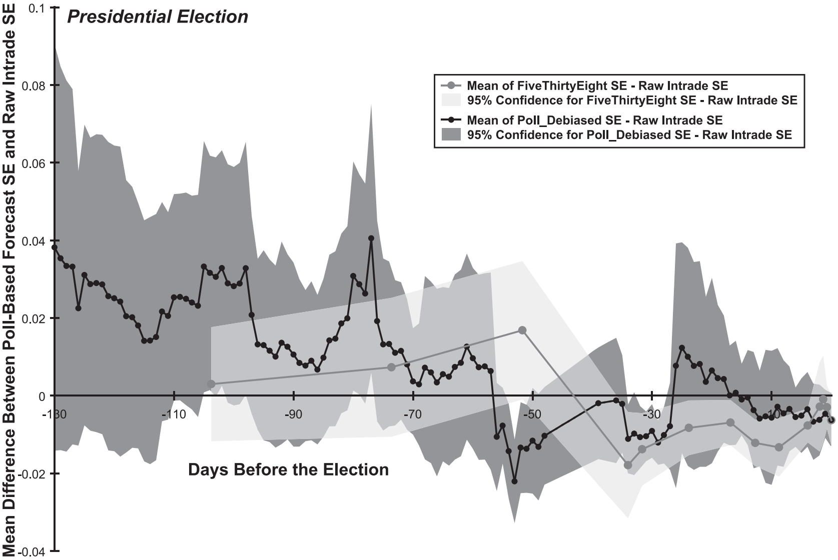

Figure A3. Mean of Poll-Based Forecast's Square Error—Raw Intrade's Square Error with 95% Confidence Interval
NOTE.—Each point plots the difference in MSE arising from the forecasts issued at that point in time for the 50 Presidential Electoral College races in the sample.

Supplementary Data

Supplementary data are available online at http://poq.oxfordjournals.org/.

References

Berg, Joyce, Robert Forsythe, Forrest Nelson, and Thomas Rietz. 2001. "Results from a Dozen Years of Election Futures Market Research." In *Handbook of Experimental Economic Results*, eds. Charles Plott and Vernon Smith. Amsterdam, The Netherlands: Elsevier.

Campbell, James E. 2000. *The American Campaign*. College Station: Texas A&M University Press.

Erikson, Robert S., and Christopher Wlezien. 2002. "The Timeline of Presidential Election Campaigns." *The Journal of Politics* 64(4):969–93.

———. 2008. "Are Political Markets Really Superior to Polls as Election Predictors?" *Public Opinion Quarterly* 72:190–21.

Fair, Ray, and Robert Shiller. 1989. "The Informational Content of ex-Ante Forecasts." *Review of Economics and Statistics* 71(2):325–31.

———. 1990. "Comparing Information in Forecasts From Econometric Models" *American Economic Review* 80(3):375–89.

Leigh, Andrew, Justin Wolfers, and Eric Zitzewitz. 2007. "Is There a Favorite-Longshot Bias in Election Markets?" Preliminary version presented at the 2007 UC Riverside Conference on Prediction Markets, Riverside, CA, USA.

Manski, Charles F. 2005. “Interpreting the Predictions of Prediction Markets.” NBER Working Paper No. 10359.

Snowberg, Erik, Justin Wolfers, and Eric Zitzewitz. 2007. “Party Influence in Congress and the Economy.” *Quarterly Journal of Political Science* 2:277–86.

Wolfers, Justin, and Eric Zitzewitz. 2004. “Prediction Markets.” *Journal of Economic Perspectives* 18(2):107–26.

———. 2007. “Interpreting Prediction Market Prices as Probabilities.” NBER Working Paper No. 12200.