Spring 5-16-2011

Forecasting: Expectations, Intentions, and Confidence

David Michael Rothschild
University of Pennsylvania's Wharton School, rothscdm@wharton.upenn.edu

Forecasting: Expectations, Intentions, and Confidence

Abstract
All three articles in my dissertation gather information from individuals, analyze it, and aggregate that information into forecasts of upcoming events. The motivation is to make forecasts more efficient (accurate and timely), more versatile (provide the most useful information for each stakeholder), and more economically efficient (equally or more efficient and versatile for less time and/or money). The first article looks at prediction markets and polls and concludes that prediction market-based forecasts are more efficient. The two methods, polling versus prediction markets, vary in four key ways: sample selection (a random sample of representative group versus a self-selected group), question type (intention versus expectation), aggregation method (average versus weighted by money, a proxy for confidence), and incentive (not incentive compatible *versus* incentive compatible). The second article isolates the second aspect of that list by comparing the efficiency of forecasts created by polling the respondents on their expectations versus intentions. Expectation-based forecasts are more efficient, even using non-random samples for the expectation. Asking the expectation question to one respondent is the equivalent of asking several respondents the intention question. Further, the expectation question helps adjust the sample to be more representative of the target group. The third article tests a new interactive web-based interface that captures both "best estimate" point-estimates and probability distributions from non-experts. In contrast to standard methods of directly asking respondents to state their confidence, using my method, which induces the respondents to reveal confidence, there is a sizable and statically significant positive relationship between confidence and the accuracy of individual-level expectations. This positive correlation between confidence and accuracy can be utilized to create confidence-weighted aggregated forecasts that are more efficient than the standard "consensus forecasts."

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Applied Economics

First Advisor
Justin Wolfers

Keywords
Forecasting, Polling, Information Aggregation, Belief Heterogeneity, Expectations, Confidence

Subject Categories
Behavioral Economics | Political Economy

FORECASTING:
EXPECTATIONS, INTENTIONS, AND CONFIDENCE

David Michael Rothschild

A DISSERATION
in
Business and Public Policy
For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania
in
Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
2011

Supervisor of Dissertation:

Justin Wolfers, Associate Professor, Business and Public Policy

Graduate Group Chairperson:

Eric Bradlow, Professor, Marketing, Statistics, and Education

Dissertation Committee:
Alex Gelber, Assistant Professor, Business and Public Policy
Sunshine Hillygus, Associate Professor, Political Science (Duke University)
Robert Inman, Professor, Finance
Uri Simonsohn, Associate Professor, Operations and Information Management
Justin Wolfers, Associate Professor, Business and Public Policy

Dedication:

Professor Wendy Heltzer, PhD
Judge James Rothschild Jr.
My cohort (It really helps to have a bunch of smart people around all of time!)
My friends and family (Thanks Mom!)

Advisor:

Justin Wolfers

A special thank you to Sunshine Hillygus for her tireless help on the first article and to Alex Gelber for his guidance on the third article. And to all of those people who commented on drafts of all, or some, of these articles: Greg Bilton, Bob Erikson, Sarah Iosifescu, Adam Isen, Andrew Leonard, Neil Malhotra, Marc Meredith, Katherine Milkman, Andrew Paciorek, Janet Pack, Betty Rothschild, Dick Rothschild, Uri Simonsohn, Erik Snowberg, Richard Thaler, Jeremy Tobacman, Christopher Wlezien, and Eric Zitzewitz. Finally, to all of the seminar audiences from BPUB 900 through the job market who provided me comments not just on these final three articles, but ones that never made it to final dissertation … Thank You!

ABSTRACT

FORECASTING: EXPECTATIONS, INTENTIONS, AND CONFIDENCE

David Michael Rothschild
Advisor: Justin Wolfers

All three articles in my dissertation gather information from individuals, analyze it, and aggregate that information into forecasts of upcoming events. The motivation is to make forecasts more efficient (accurate and timely), more versatile (provide the most useful information for each stakeholder), and more economically efficient (equally or more efficient and versatile for less time and/or money). The first article looks at prediction markets and polls and concludes that prediction market-based forecasts are more efficient. The two methods, polling versus prediction markets, vary in four key ways: sample selection (a random sample of representative group versus a self-selected group), question type (intention versus expectation), aggregation method (average versus weighted by money, a proxy for confidence), and incentive (not incentive compatible *versus* incentive compatible). The second article isolates the second aspect of that list by comparing the efficiency of forecasts created by polling the respondents on their expectations versus intentions. Expectation-based forecasts are more efficient, even using non-random samples for the expectation. Asking the expectation question to one respondent is the equivalent of asking several respondents the intention question. Further, the expectation question helps adjust the sample to be more representative of

the target group. The third article tests a new interactive web-based interface that captures both "best estimate" point-estimates and probability distributions from non-experts. In contrast to standard methods of directly asking respondents to state their confidence, using my method, which induces the respondents to reveal confidence, there is a sizable and statically significant positive relationship between confidence and the accuracy of individual-level expectations. This positive correlation between confidence and accuracy can be utilized to create confidence-weighted aggregated forecasts that are more efficient than the standard "consensus forecasts."

Table of Contents

Article 1
Page 1 … Forecasting Elections: Comparing Prediction Markets, Polls and their Biases

Article 2
Page 30 … Forecasting Elections: Voter Intentions versus Expectations

Article 3
Page 76 … Expectations: Point-Estimates, Probability Distributions, Confidence, and Forecasts

List of Tables

Article 1

p23: **Table 1.** Coefficients from Probit of Winner on Forecasts, where the dependent variable is $I(IncumbentWin)_r$ (Individual Regressions)

p24: **Table 2.** Coefficients from Probit of Winner on Forecasts, where the dependent variable is $I(IncumbentWin)_r$ (Joint Regressions)

Article 2

p35: **Table 1.** Forecasting the Winner of the Presidential Races

p39: **Table 2.** Forecasting the Presidential Election, by State

p47: **Table 3**. Comparing the Accuracy of Naïve Forecasts of Vote Shares

p58: **Table 4.** Comparing the Accuracy of Efficient Forecasts of Vote Share

p62: **Table 5.** Comparing the Accuracy of Efficient Probabilistic Forecasts

p63: **Table 6.** Comparing the Accuracy of Efficient Forecasts for the 2008

p66: **Table 7.** Comparing the Accuracy of Efficient Forecasts of Vote Shares from Biased Samples

p67: **Table 8.** Comparing the Accuracy of Efficient Probabilistic Forecasts

p68: **Table 9.** Comparing the Accuracy from Secondary Dataset

p71: **Table 10.** Relationship between Voter Intention, the Winner, and Expectation

Article 3

p85: **Table 1.** Coefficients of Variation of Individual-Level Probability Distributions and Coefficients of Variation of Point-Estimates

p88: **Table 2.** Correlations Between Confidence and Accuracy of Point-Estimate

p90: **Table 3.** Individual-Level Point-Estimates

p91: **Table 4.** Comparing the Point Accuracy of the Individual-Level Expectations

p93: **Table 5.** Comparing the Point Accuracy of the Most Promising Forecasts

p94: **Table 6.** Comparing the Point Accuracy of the Most Promising Forecasts

List of Figures

Article 1

p11: **Figure 1.** Probability of Victory in the National Popular Vote for the Incumbent Party Candidate in the 2008 Presidential Election

p14: **Figure 2.** Distribution of Probabilities of Victory for Winning Candidates

p17: **Figure 3.** Probability of Victory for the Winning Candidate in Select Races

p20: **Figure 4.** Mean of Poll-Based Forecast's Square Error - Debiased Intrade's Square Error with 95% Confidence Interval (Presidential Electoral College Races)

p21: **Figure 5.** Mean of Poll-Based Forecast's Square Error - Debiased Intrade's Square Error with 95% Confidence Interval (Senatorial Elections)

p27: **Figure A1.** Comparison of Pageviews for FiveThirtyEight.com and Intrade.com During the 2008 Election Cycle.

p27: **Figure A2.** Probability of Victory in the National Popular Vote for the Incumbent Party Candidate in the 2008 Presidential Election

p28: **Figure A3.** Mean of Poll-Based Forecast's Square Error - Raw Intrade's Square Error with 95% Confidence Interval (Presidential Electoral College Elections)

Article 2

p41: **Figure 1**. Naïve Voter Intention Forecast and Actual Vote Share

p42: **Figure 2.** Voter Expectation and Actual Vote Share

p45: **Figure 3.** Naïve Expectation-Based Forecast and Actual Vote Share

p49: **Figure 4.** Sample Size and Forecast Errors in the Intent Poll

p53: **Figure 5.** Efficient Intention-Based Forecasts and Actual Vote Share

p57: **Figure 6.** Efficient Expectation-Based Forecast and Election Outcomes

p60: **Figure 7.** Sample Size and Forecast Errors in the Forecasts of Vote Share

p73: **Figure 8.** Relationship between Intention, Expectation, and Actual Vote Share

Article 3

p88: **Figure 1.** Correlations Between Confidence Derived From Probability Distributions and Accuracy of Point-Estimate

Forecasting Elections:
Comparing Prediction Markets, Polls and their Biases

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.

I. Introduction

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 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

[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 point).

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

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]

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

---

[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.

[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 elections 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 2x that of Intrade's to over 7.5x; on Election Day FiveThirtyEight had an astonishing five million page views. See Appendix, Figure A1, for chart of page views. Data is from http://www.alexa.com/siteinfo/fivethirtyeight.com and

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).

II. 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).

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

---

http://www.nytimes.com/2008/11/10/business/media/10silver.html?scp=5&sq=Stephanie%20Clifford&st=cse.

into prices. Wolfers and Zitzewitz (2007) show that in addition to non-risk 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).

---

[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.

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 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.

III. 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

[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.

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 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: $V_{yr} = \alpha + \beta_{yr} + e_{yr}$, 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: $\widehat{V_{2008,T}} = \alpha_T + \beta_T P_{2008,T}$; the alpha corrects for the anti-incumbency bias and the beta corrects for reversion to the mean. For Presidential races, FiveThirtyEight projects the snapshot

[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.
[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.

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: $Pr = \Phi\left(\frac{\widehat{V_{2008,T}}}{RMSE_T + V(\beta_T)(P_{2008,T} - \overline{P_T})}\right)$. For Poll_Debiased I use maximum likelihood to determine the optimal sigma (σ_T) for each day: $Pr = \Phi\left(\frac{\widehat{V_{2008,T}}}{\sigma_T}\right)$.[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

[12] For all of Poll_Debiased's parameters I use $\pm$ 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.

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 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): $Pr = \Phi\left(1.64 * \Phi^{-1}(price)\right)$.[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 nineteen 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

---

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

[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.

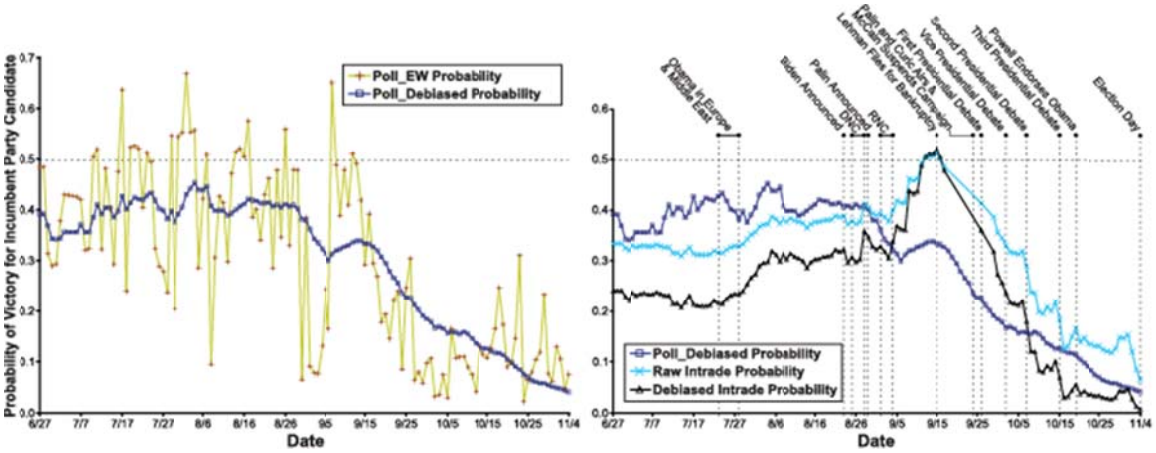
[18] Unfortunately, there are ten random days where I do not have Intrade data; those days are dropped from the direct comparisons with Poll_EW and Poll_Debiased.

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 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]

[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.

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



Notes: 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.

IV. 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. 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.

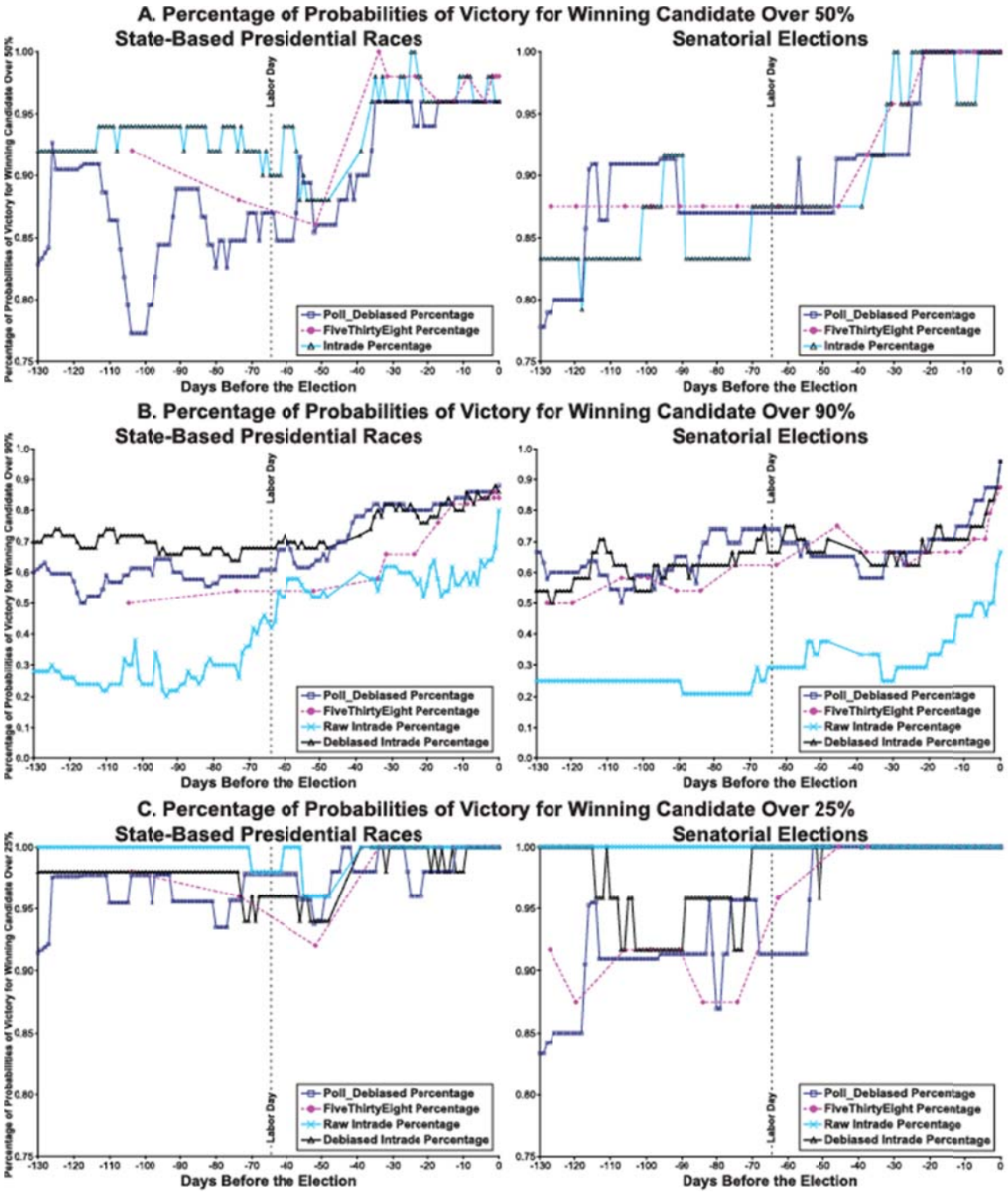
In probabilities above 90 percent FiveThirtyEight is a cautious predictor, moving toward these more certain probabilities late in the cycle; Poll_Debiased and Debiased Intrade are similar to each other, with Debiased Intrade showing slightly more confident probabilities early in the Presidential cycle (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.

[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.

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 post-convention 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.

Figure 2: Distribution of Probabilities of Victory for Winning Candidates



Notes: 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 50 Presidential Electoral College races in the sample; on the right side, it is from the 24 Senatorial elections.

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 candidates lost when they were predicted to win with >90% 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 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

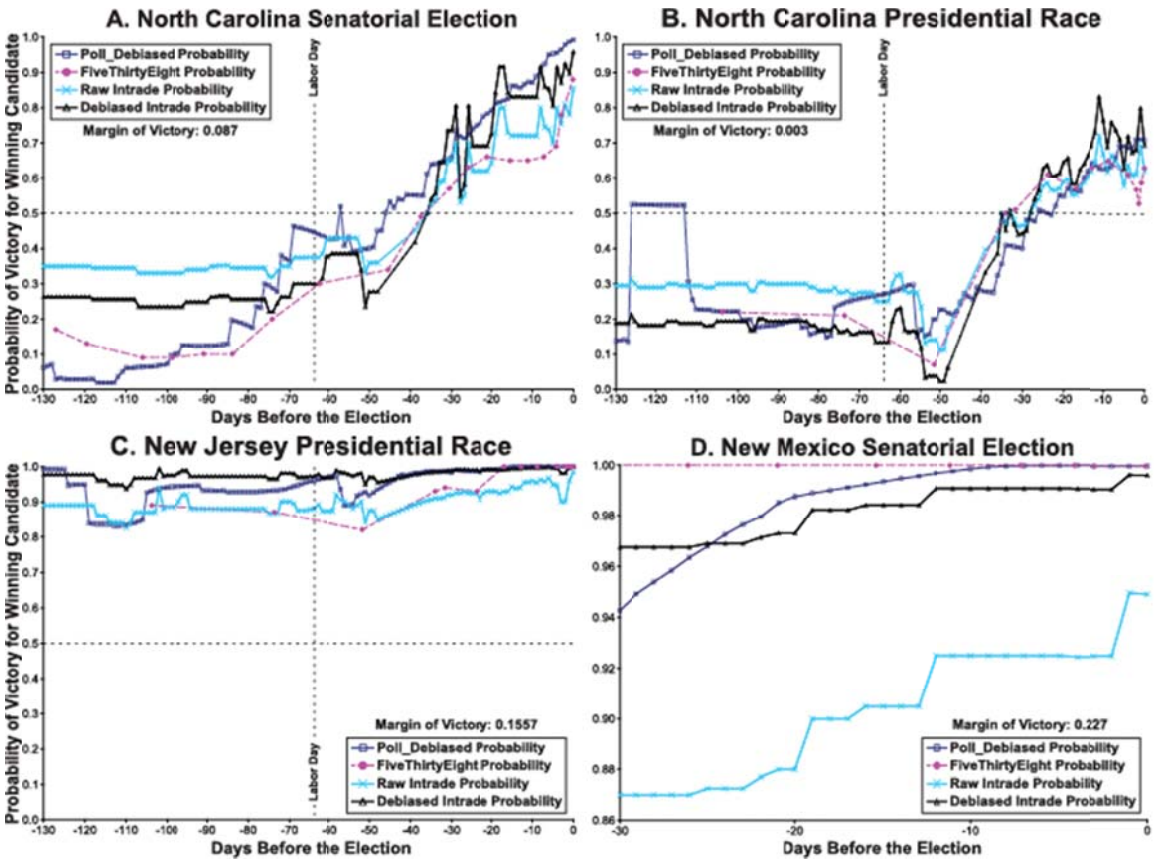
[21] Any finding in this paper relating to the 90 percent line is robust to nearby probabilities.
[22] FiveThirtyEight and Debiased Intrade have nineteen observations, out of over 9,500 total observations, where they give >90 percent probability of victory to the eventual losing candidate. These observations 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 3 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.

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]

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



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 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

[24] Please see the supplementary data online (figure A4) for charts 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.

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).

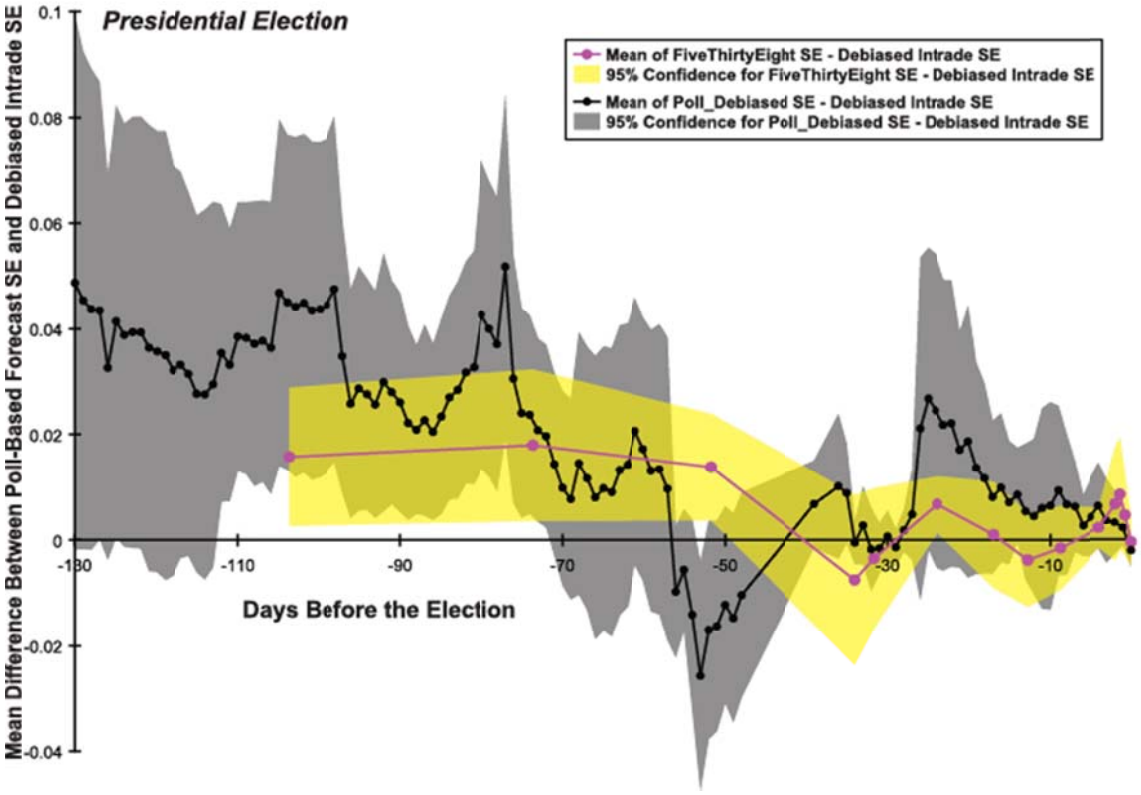
For presidential races, as shown in figure 4, Debiased Intrade has a statistically significant smaller mean square error relative to the poll-based forecasts until mid-September and then continues to have smaller errors until the end of the cycle.[26] The figure charts the mean difference in square error 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

[25] Please see 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.

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 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.

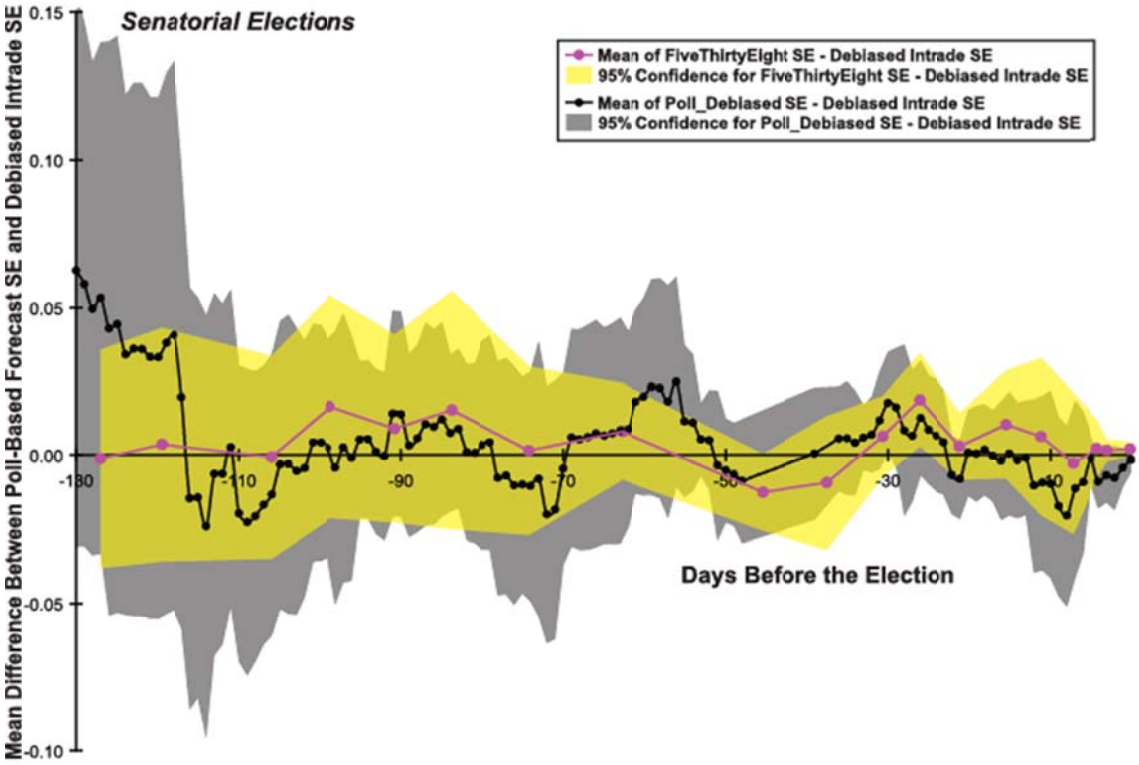
Figure 4: Mean of Poll-Based Forecast's Square Error – Debiased Intrade's Square Error with 95% Confidence Interval



Notes: Each point plots the difference in MSE arising from the forecasts issued at that point in time for the 50 Presidential Electoral Colleges races in the sample.

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.

Figure 5: Mean of Poll-Based Forecast's Square Error – Debiased Intrade's Square Error with 95% Confidence Interval



Notes: Each point plots the difference in MSE arising from the forecasts issued at that point in time for the 24 Senatorial elections in the sample.

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 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:

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

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 non-incumbent candidate for all of their 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 non-incumbent. 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 non-incumbent.

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.

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

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

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

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 unique information from each other:

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

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.

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

Notes: (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.

V. 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.

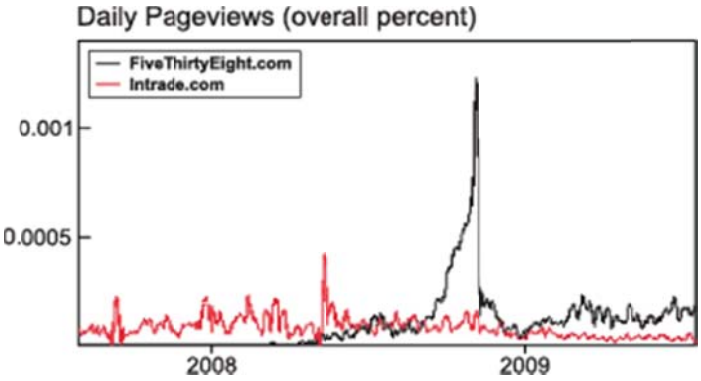
As to information, the raw information used by the poll-based forecasts is public and hence should be in the information set of Intrade investors. Beyond this, prediction markets aggregate dispersed and unpublished information (i.e., a brewing scandal may be known to a few investors before the general 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.

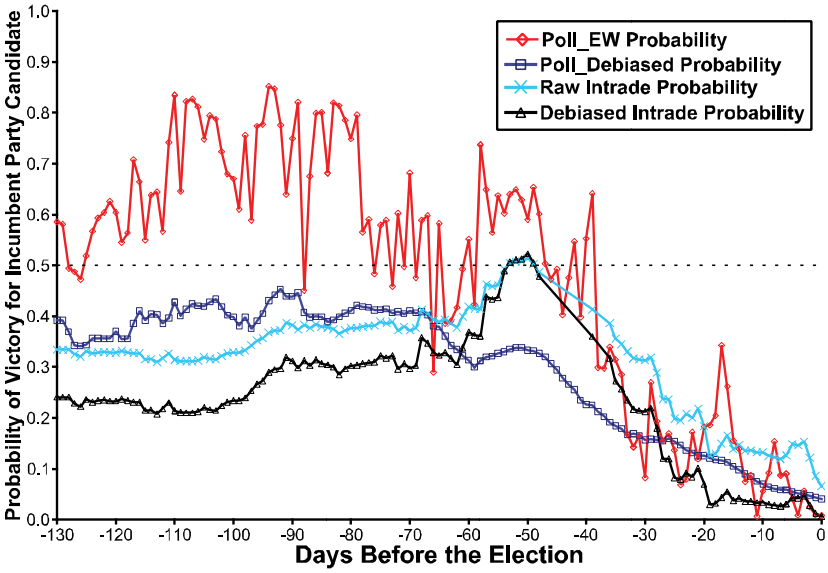
VII. Appendix

Figure A1: Comparison of Pageviews for FiveThirtyEight.com and Intrade.com During the 2008 Election Cycle



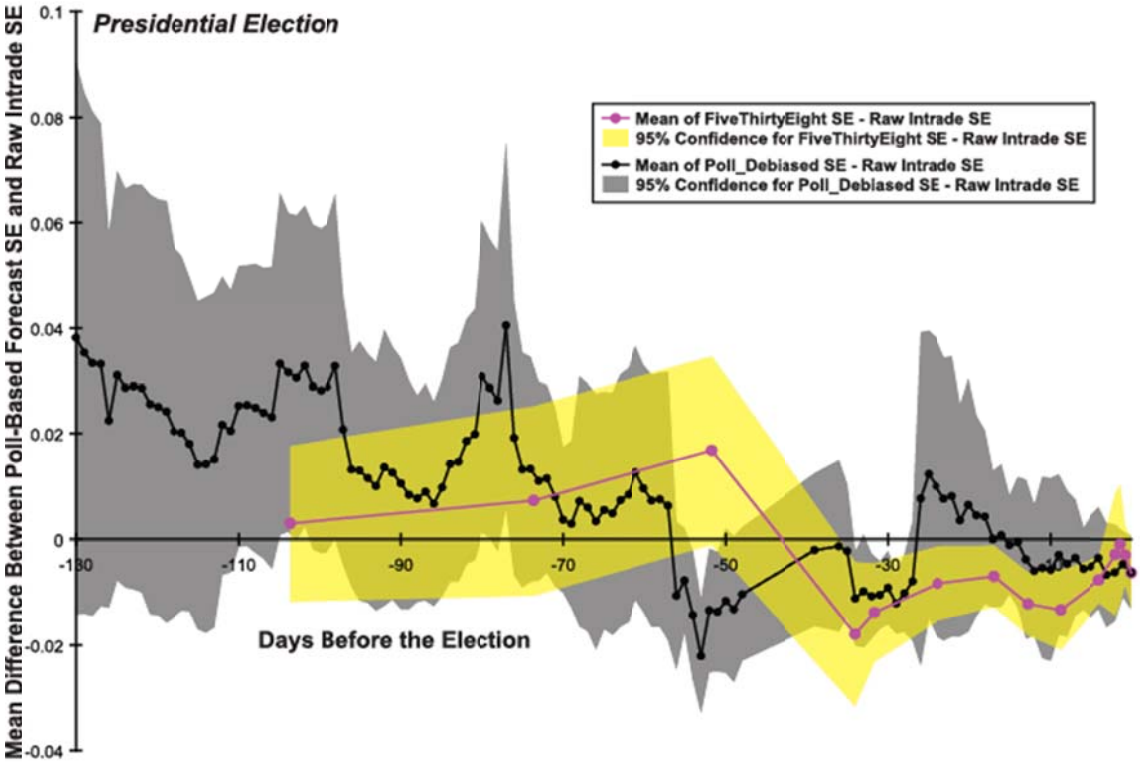
Notes: FiveThirtyEight.com launched in March of 2008. Data is from Alexa.com

Figure A2: Probability of Victory in the National Popular Vote for the Incumbent Party Candidate in the 2008 Presidential Election



Notes: This figure differs from Figure 1 because instead of using data from the 2000 and 2004 state by state Presidential Electoral Colleges 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 – Debiased Intrade's Square Error with 95% Confidence Interval



Notes: Each point plots the difference in MSE arising from the forecasts issued at that point in time for the 50 Presidential Electoral Colleges races in the sample.

VIII. 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*, 2:277-286.

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

Forecasting Elections:
Voter Intentions versus Expectations

(Article is co-authored with Justin Wolfers)

Abstract

In this paper, we explore the value of an underutilized political polling question: who do you think will win the upcoming election? We demonstrate that this expectation question points to the winning candidate more often than the standard political polling question of voter intention: if the election were held today, who would you vote for? Further, the results of the expectation question translate into more accurate forecasts of the vote share and probability of victory than the ubiquitous intent question. This result holds, even if we generate forecasts with the expectations of only Democratic voters or only Republican voters and compare those forecasts to forecasts created with the full sample of intentions. Our structural interpretation of the expectation question shows that every response is equivalent to a multi-person poll of intention; the power of the response is that it provides information about the respondent's intent, as well as the intent of her friends and family. This paper has far reaching implications for all disciplines that use polling.

I. Introduction

Since the advent of scientific polling in the 1930s, political pollsters have asked people whom they intend to vote for; occasionally, they have also asked who they think will win. Our task in this paper is long overdue: we ask which of these questions yields more accurate forecasts. That is, we contrast the predictive power of the questions probing voter *intention* with questions probing *expectation*. Judging by the attention paid by pollsters, the press, and campaigns, the conventional wisdom appears to be that polls of voter intention are more accurate than polls of voter expectation.

Yet there are good reasons to believe that the expectation question is more informative. Survey respondents may possess much more information about the upcoming political race than the voting intention question allows them to answer. At a minimum, they know their own current voting intention, so the information set feeding into their expectation will be at least as rich as that captured by the voting intention question. But beyond this, they may also have information about the current voting intentions, both preference and probability of voting, of their friends and family. So too, they have some sense of the likelihood that today's expressed intention will be changed before it ultimately becomes an Election Day vote. Our research is motivated by idea that the richer information embedded in this expectation data may yield more accurate forecasts.

We find robust evidence that expectation-based forecasts yield more accurate predictions of election outcomes. By comparing the performance of these two questions only when they are asked in exactly the same survey, we effectively difference out the influence of other factors. Our primary dataset consists of all of the Presidential Electoral College races from 1952 to 2008, where both the intention and expectation question are asked. In 268 of the 345 polls, either both the intention and expectation question point to the winner or neither does. But in the 77 cases in which one points to the winner and the other does not, the expectation question points to the winner 60 times, while the intention question points to the winner only 17 times. That is, 78% of the time that these two approaches disagree, the expectation data is correct. We can also assess the relative accuracy of the two methods by assessing the extent to which each can be informative in forecasting the vote share and the probability of victory; we find that relying on voter expectation rather than intention data yield substantial and statistically significant increases in forecasting accuracy. Our findings remain robust to correcting for an array of known biases in voter intention data.

The better performance of forecasts created with expectation question, versus intention question, data varies somewhat, depending on the specific context. The expectation-based question is particularly valuable when small samples are involved. The intuition for this result comes from a simple thought experiment. In our primary dataset, we have 13,208 individual respondents providing their intention and expectation in 345 different races; 58.0% of respondents intend to vote for the winning candidate, while 68.5% expect that candidate to win. Thus, if we survey only one voter, expectation will outperform intention 10.5% of the time. It is unclear ex-ante which type of question relatively benefits from increases in the days before the election, although both are less accurate as less information is available.

One strand of literature this paper addresses is the emerging documentation that prediction markets tend to yield more accurate forecasts than polls (Wolfers and Zitzewitz, 2004; Berg, Nelson and Rietz, 2008). More recently, Rothschild (2009) has updated these findings in light of the 2008 Presidential and Senate races, showing that forecasts based on prediction markets yielded systematically more accurate forecasts of the likelihood of Obama winning each state than did the forecasts based on aggregated intention polls compiled by the website FiveThirtyEight.com and another more transparent intention poll-based forecast created by the author. One hypothesis for this superior performance is that prediction markets—by asking traders to bet on outcomes—effectively ask a different question, eliciting the expectations rather than intentions of participants. If correct, this suggests that much of the accuracy of prediction markets could be obtained simply by polling voters on their expectations, rather than intentions.

These results also speak to yet another strand of research, the historical question about the value of scientific polling and representative samples (Robinson, 1937). Begun prior to the advent of scientific polling and renewed most recently with the rise of cellphones as well as use of online survey panels, this debate is again of contemporary

significance. Surveys of voting intentions rely heavily on polling representative cross-sections of the electorate. By contrast, as we demonstrate in this paper, surveys of voter expectation can still be quite accurate, even when drawn from non-representative samples. Again, the logic of this claim comes from the difference between asking about expectations, which should not systematically differ across demographic groups, and asking about intentions, which should. Again, the connection to prediction markets is useful, as Berg and Rietz (2006) shows that these have yielded accurate forecasts, despite drawing from an unrepresentative pool of overwhelmingly white, male, highly educated, high income, self-selected traders.

While direct voter expectation questions about electoral outcomes have been virtually ignored by political forecasters, they have received some interest from psychologists. In particular, Granberg and Brent (1983) document wishful thinking, in which people's expectation about what will occur is positively correlated with what they want to happen. Thus, people who intend to vote Republican are also more likely to predict a Republican victory. This same correlation is also consistent with voters preferring the candidate they think will win, as in bandwagon effects, or gaining utility from being optimistic. We re-interpret this correlation through a rational lens, in which the respondents know their own voting intention with certainty and have knowledge about the voting intentions of their friends and family. Insights from this structural interpretation of the data both explain the power of the expectation data and, by revealing the relationship between intention and expectation, may help us identify even more efficient translations of these two sets of raw data into the underlying values of the election.

More accurate forecasts will provide researchers a tool for capturing the impact of campaigns on elections; this is currently a difficult question to address, as there is great variation in both the slope and values of currently utilized forecasts of elections. Our method will also allow for forecasts of campaigns that are currently too difficult or

costly to poll. Beyond understanding the effect of campaigns on elections, the findings in this paper are important as forecasts affect races, as resources are allocated by the progress of the campaign (Mutz, 1995), and voters themselves act strategically in certain contexts (Irwin and Holsteyn, 2002). Political forecasts also have consequences beyond politics, as individual companies and markets react to the probability of different outcomes (Imai and Shelton, 2010).

We believe that our findings have substantial applicability in other forecasting contexts. Market researchers ask variants of the voter intention question in an array of contexts; as they read this paper, they can seamlessly substitute a product launch in place of an election, the preference for one product over another in place of voter intention, and the consumer expectation of sales for one product over another in place of voter expectation. Likewise, indices of consumer confidence are partly based on the stated purchasing intentions of consumers, rather than their expectations about the purchase conditions for their community. The same insight that motivated our study—that people also have information on the plans of others—is also likely relevant in these other contexts. Thus, it seems plausible that other types of research may also benefit from paying greater attention to people's expectations than to their intentions.

In Section II, we describe our first cut of the data, illustrating the relative success of the two approaches to predicting the winner of elections. In Section III, we create a naïve translation of the raw data into forecasts of vote share. In Section IV, we generate a more efficient translation of the raw data into forecasts of vote share. In Section V, we determine the accuracy of the polls in creating probabilities of victory. In Section VI, we test our forecasts with 2008 data. In Section VII, we examine the accuracy of expectation-based forecasts produced with non-random samples of respondents. In Section VIII, we assess the methods derived in this paper from the primary data source, on a secondary data source. In Section IX, we provide a structural interpretation of the response to the expectation question.

II. Simple Forecasting of the Winner

Our primary dataset consists of the American National Election Studies (ANES) cumulative data file. In particular, we are interested in responses to two questions:

Voter Intention: *Who do you think you will vote for in the election for President?*

Voter Expectation: *Who do you think will be elected President in November?*

These questions are typically asked one month prior to the election. Throughout this paper, we treat elections as two-party races, and so discard responses involving professed intention to vote for or expectation of victory for third-party candidates. In order to keep the sample sizes comparable, we only keep respondents with valid responses to both the intention and expectation questions and adjust the individual response with the ANES provided weights. When we describe the "winner" of an election, we are thinking about the outcome that most interests forecasters, which is who takes office (and so we describe George W. Bush as the winner of the 2000 election, despite his losing the popular vote).

At the national level, both questions have been asked since 1952, and to give a sense of the basic patterns, we summarize these data in Table 1.

Table 1: Forecasting the Winner of the Presidential Races

Year	Race	%Expect the winner	%Intended to vote for winner	%Reported voting for winner	Actual result: % voting for winner	N
1952	Eisenhower beat Stevenson	56.0%	56.0%	58.6%	55.4%	1,135
1956	Eisenhower beat Stevenson	76.4%	59.2%	60.6%	57.8%	1,161
1960	Kennedy beat Nixon	45.0%	45.0%	48.4%	50.1%	716
1964	Johnson beat Goldwater	91.0%	74.1%	71.3%	61.3%	1,087

1968	Nixon beat Humphrey	71.2%	56.0%	55.5%	50.4%	844
1972	Nixon beat McGovern	92.5%	69.7%	68.7%	61.8%	1,800
1976	Carter beat Ford	52.6%	51.4%	50.3%	51.1%	1,320
1980	Reagan beat Carter	46.3%	49.5%	56.5%	55.3%	870
1984	Reagan beat Mondale	87.9%	59.8%	59.9%	59.2%	1,582
1988	GHW Bush beat Dukakis	72.3%	53.1%	55.3%	53.9%	1,343
1992	Clinton beat GHW Bush	65.2%	60.8%	61.5%	53.5%	1,541
1996	Clinton beat Dole	89.6%	63.8%	60.1%	54.7%	1,274
2000	GW Bush beat Gore	47.4%	45.7%	47.0%	49.7%	1,245
2004	GW Bush beat Kerry	67.9%	49.2%	51.6%	51.2%	921
2008	Obama beat McCain	65.7%	56.6%	56.5%	53.7%	1,632
	Simple Average:	**68.5%**	**56.7%**	**54.6%**	**57.5%**	**18,471**

Notes: Table summarizes authors' calculations, based on data from the American National Election Studies, 1952-2008. Sample restricted to respondents whose responses to both the expectation and intention questions listed the two major candidates.

Each method can be used to generate a forecast of the most likely winner, and so we begin by assessing how often the majority response to each question correctly picked the winner. The first column with data on Table 1 shows that the winning Presidential candidate was expected to win by a majority of respondents in 12 of the 15 elections, missing Kennedy's narrow victory in 1960, Reagan's election in 1980, and G.W. Bush's controversial win in 2000. The more standard voter intention question performed similarly, correctly picking the winning candidate in one fewer election. The only election in which the two approaches pointed to different candidates was 2004, in which a majority of respondents correctly expected that Bush would win, while a majority intended to vote for Kerry. So far we have been analyzing data from the pre-election interviews. In the third column we summarize data from post-election interviews which also ask which candidate each respondent ultimately voted for. The data in this column

reveal the influence of sampling error, as a majority of the people sampled in 1960 and 2000 ultimately did vote for the losing candidate.

The last line of this table summarizes, and on average, 68.5% of all voters correctly expected the winner of the Presidential election, while 56.7% intended to vote for the winner. These averages give a hint as to the better performance of expectation-based forecasts. Taken literally, they say that if one forecasted election outcomes based on a random sample of one person in each election, asking about voter expectations would predict the winner 68.5% of the time, compared to 56.7%, when asking about voter intentions. More generally, in small polls, sampling error likely plays a larger role in determining whether a majority of respondents intend to vote for the election winner, than in whether they correctly forecast the winner. We will develop this insight at much greater length, in section IV.

The analysis in Table 1 does not permit strong conclusions, and indeed, it highlights two important analytic difficulties. First, we have very few national Presidential elections, and so the data will permit only noisy inferences. Second, our outcome measure—asking whether a method correctly forecasted the winner—is a very coarse measure of the forecasting ability of either approach to polling. Thus, we will proceed in two directions. First, we will exploit a much larger number of elections by analyzing data from the same surveys on who respondents expect to win the Electoral College votes of their state. And second, we will proceed to analyzing the forecasting performance of each approach against other measures: their ability to match the two-party preferred vote share and their forecast of the probability of victory.

We begin with the state-by-state analysis, analyzing responses to the state-specific voter expectation question:

Voter Expectation (state level): *How about here in [state]. Which candidate for President do you think will carry this state?*

We compare responses to this question with the voter intention question described above. Before presenting the data, there are four limitations of these data worth noting. First, the ANES does not survey people in every state, and so in each wave, around 35 states are represented. Second, this question was not asked in the 1956-68 and 2000 election waves. We do not expect either of these issues to bias our results toward favoring either intention- or expectation-based forecasts. Third, the sample sizes in each state can be small. Across each of these state elections, the average sample size is only 38 respondents, and the sample size in a state ranges from 1 to 246. In section IV we will see that this is an important issue, as the expectation-based forecasts are relatively stronger in small samples. Fourth, while the ANES employs an appropriate sampling frame for collecting nationally representative data, it is not the frame that one would design were one interested in estimating state-specific aggregates, as these samples typically involve no more than a few Primary Sampling Units. Despite these limitations, this data still presents an interesting laboratory for testing the relative efficacy of intention versus expectation-based polling. All told we have valid ANES data from 10 election cycles (1952, 1972-1996, and 2004-2008), and in each cycle, we have data from between 28 and 40 states, for a final sample size of 345 races.[28]

Table 2 summarizes the performance of our two questions at forecasting the winning Presidential candidate in each state. Again, we use a very coarse performance metric, simply scoring the proportion of races in which the candidate who won a majority in the relevant poll ultimately won in that state; the voter expectation question is the first column with data and the usual voter intention question is the second column with data. (When a poll yields a fifty-fifty split, we score it as half of a correct call.) All told, the voter expectation question predicted the winner in 279 of these 345 races, compared with 239 correct calls for the voter intention question. A simple difference in

[28] Only the 311 pre-2008 elections are used in Sections III, IV, and V to test and calibrate our models. Section VI then runs the model, with coefficients created with pre-2008 data, on 2008 data.

proportions test reveals that these differences are clearly statistically significant (z=3.63). Of course the forecast errors of each approach may be correlated across states within an election cycle, and so a more conservative approach would note that the voter expectation question outperformed the voter intent question in 8 of the 10 election cycles and tied in 2008. The difference in that tenth cycle (1972) was that Nixon won a tight race in Minnesota. More to the point, this difference in forecasting performance is large.

Table 2: Forecasting the Presidential Election, by State

Year	Proportion of states where the winning candidate was correctly predicted by a majority of respondents to:		Number of states surveyed
	Expectation question	Intention question	
1952	74.3%	58.6%	35
1972	97.4%	100%	38
1976	80.3%	77.6%	38
1980	57.7%	41.0%	39
1984	86.7%	68.3%	30
1988	88.3%	56.7%	30
1992	89.4%	77.3%	33
1996	75.0%	67.5%	40
2004	89.3%	67.9%	28
2008	76.5%	76.5%	34
Totals:	**279 of 345 correct**	**239 of 345 correct**	**Difference:**
Average:	**80.9%**	**69.3%**	**11.6%*****
(Standard error)	**(3.8)**	**(5.4)**	**(3.2)**

Notes: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10%, respectively. (Standard errors in parentheses are clustered by year).

At this point, our analysis has been quite crude—only analyzing whether the favored candidate won. This approach has the virtue of transparency, but it leaves much of the variation in the data—such as variation in the winning margin—unexamined. Thus we now turn to analyzing the accuracy of the forecasted vote shares derived from both intention and expectation data. We will also add some structure to

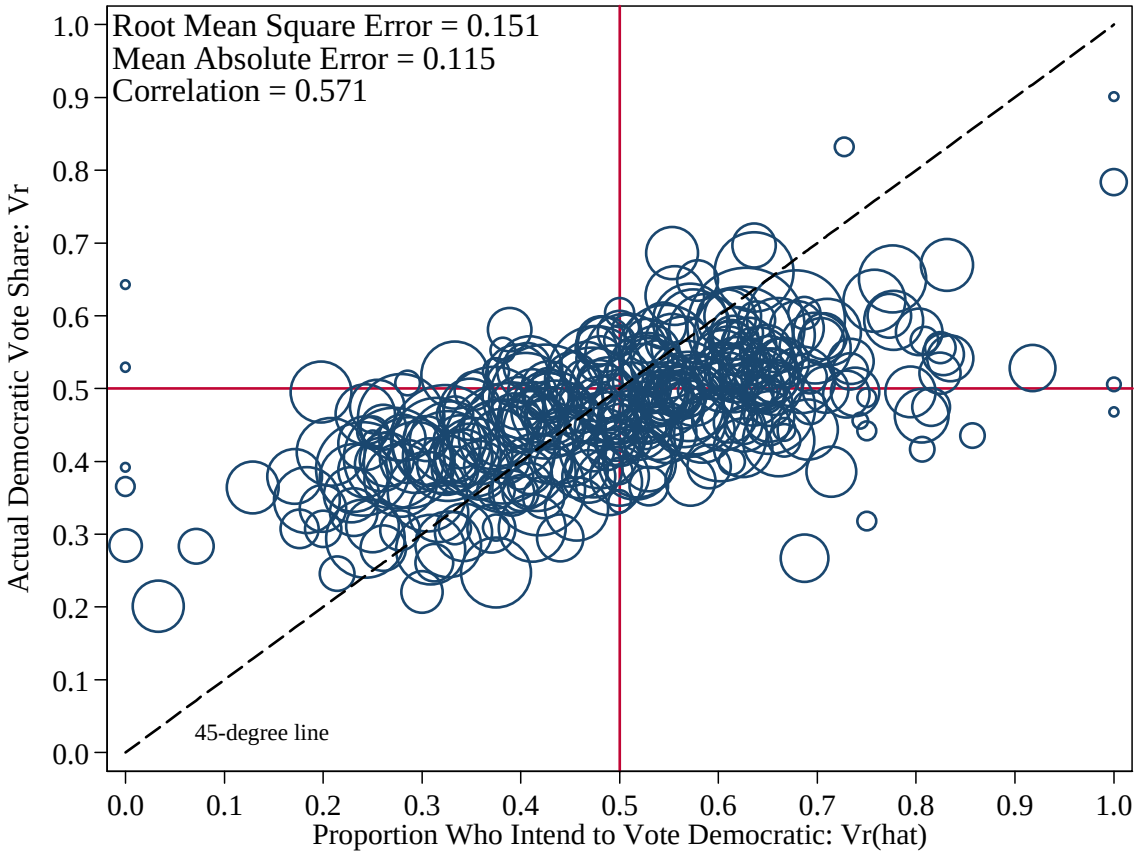
how we are thinking about these data. At this point we drop 2008 from the dataset, in order to have some "out of-sample" data to review in Section VI.

III. Simple (or Naïve) Forecasting of Vote Share

Our goal is to use the state-by-state ANES data to come up with forecasts of the two-party vote share in each of the R state×year races in our dataset. In this section, we analyze the data the way they are typically used—interpreting a poll that says that $\widehat{v_r}$ percent of *sample* respondents will vote for a candidate in state-year race r as a forecast that this candidate will win v_r percent of votes among the entire *population*. That is, we follow the norm among pollsters and make our projections as if the sample moments represent population proportions. Likewise, we interpret a poll that says that $\widehat{x_r}$ percent of sample respondents expect a candidate to win as a forecast that x_r percent of the population expect that candidate to win. While this may sound obvious, in fact raw polling data rarely represent optimal forecasts. Thus, we refer to the projections in this section as "naïve forecasts", and in section IV we will describe how our raw polling data can be adjusted to create efficient forecasts.

Our focus is on predicting each candidate's share of the two-party vote, and we begin by analyzing data on voter intention. Figure 1 plots the relationship between the actual Democratic vote share in each state-year race, and the proportion of poll respondents who plan to vote for the Democratic candidate. There are two features of these data to notice. First, election outcomes and voter expectations are clearly positively correlated—that is, these polls are informative. But second, the relationship is by no means one-for-one, and these relatively small polls of voter intention are only a noisy measure of the true vote share on Election Day.

Figure 1: Naïve Voter Intention Forecast and Actual Vote Share

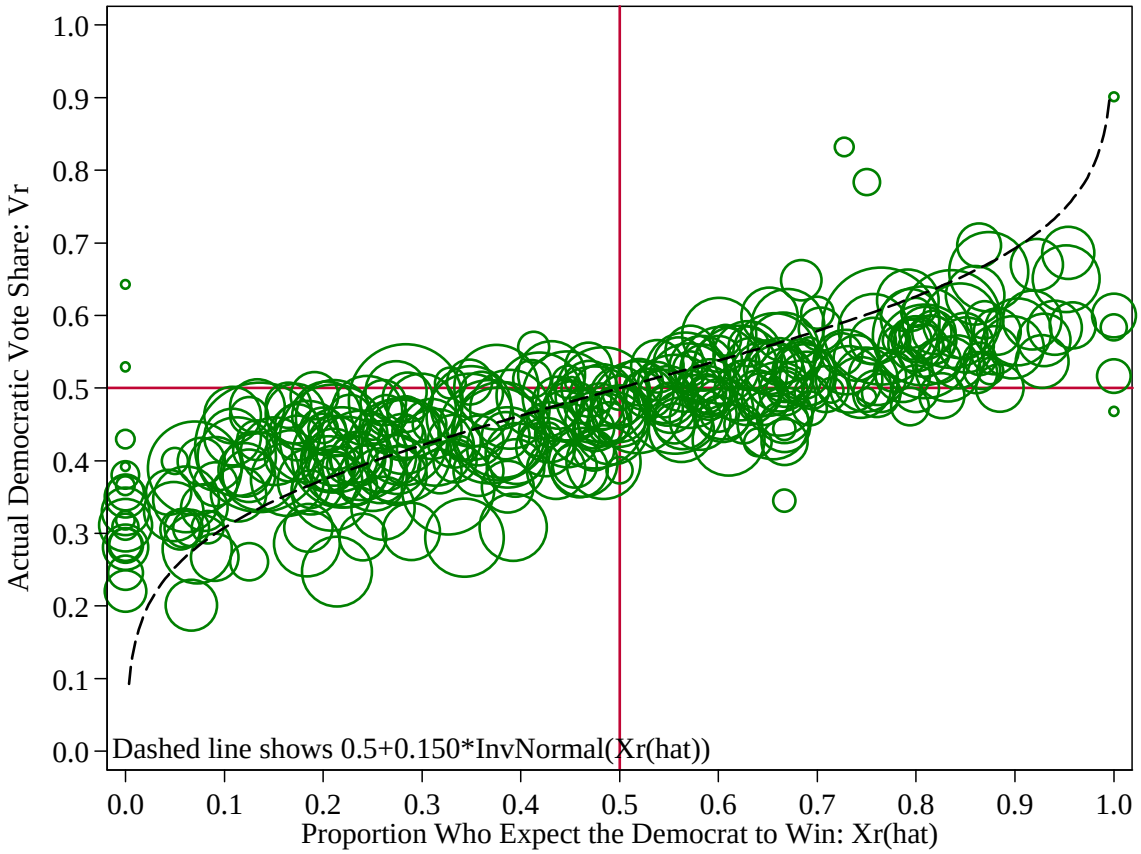


Notes: Each point shows a separate state-year cell in a Presidential Electoral College election; the size of each point is proportional to the number of survey respondents. Both voter intention and election outcomes refer to shares of the total votes cast for the two major parties. There are a total of n=311 elections, as the 2008 data is not included.

In Figure 2 we show the relationship between voter expectations—the share of voters who expect the Democrat to win that state's presidential ballot—and the vote share he actually garnered. This plot reveals that there is a close relationship between election outcomes and voter expectations, and typically the candidate who most respondents expect to win, does in fact win. That is, most of the data lie in either the Northeast or Southwest quadrants, a fact also evident in Table 2. Equally, the relationship between voter expectations and vote shares does not appear to be linear. Indeed, it should seem obvious that a statement that two-thirds of voters expect Obama to win does not—without adding further structure—immediately correspond to any

particular forecast about his likely vote share. Thus we now turn to assessing how to tease out the forecast of vote shares implicit in these data.

Figure 2: Voter Expectation and Actual Vote Share



We begin by characterizing how people respond to the question probing their expectations about the likely winner. If each poll respondent views an unbiased noisy signal, ${x^*}_r^i$ of a candidate's final vote share—where the superscript i serves as a reminder that we are analyzing individual responses, and the asterisk reminds us that this is an unobserved latent variable—then:

$$ {x^*}_r^i = v_r + \epsilon_r^i \text{ and } \epsilon_r^i \sim N(0, \sigma_\epsilon^2) \qquad [1] $$

where ϵ_r^i is an idiosyncratic error reflecting the respondent's imperfect observation.[29] We assume that this noise term is drawn from a normal distribution, and its variance is constant across both poll respondents, and across elections. In turn, if poll respondents describe themselves as expecting a specific candidate to win if this noisy signal suggests that this candidate will win at least half the vote, then we will observe voter expectations as follows:

$$x_r^i = \begin{cases} 1 & if\ x^{*i}_r = v_r + \epsilon_r^i > 0.5 \\ 0 & if\ x^{*i}_r = v_r + \epsilon_r^i < 0.5 \end{cases} \quad [2]$$

Consequently the probability that an individual respondent says that they expect a candidate to win is $\Phi\left(\frac{v_r - 0.5}{\sigma_\epsilon}\right)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. That is, equations [1] and [2] together imply that we can estimate σ_ϵ from a simple probit regression explaining whether the respondent expected a candidate to win, by a variable describing the extent by which the vote share garnered by that candidate exceeds the 50% required to win:

$$x_r^i = \frac{1}{\sigma_\epsilon}(v_r - 0.5 + \epsilon_r^i) \quad [3]$$

This regression yields an estimate of $1/\widehat{\sigma_\epsilon} = 6.661$ with a standard error allowing for within-state-year correlated errors of 0.385 (n=11,548), which implies that $\widehat{\sigma_\epsilon} = 0.150$, with a standard error of 0.0089 (estimated using the delta method). For now, we simply note that this estimate provides a link between election results, and the proportion of the population who expect the Democrat to win, x_r:

$$E[x_r] = Prob\left(v_r + \epsilon_r^i > 0.5\right) = \Phi\left(\frac{v_r - 0.5}{\sigma_\epsilon}\right) \quad [4]$$

[29] Part of this error may be due to the fact that the election is still a month away; thus ϵ^i includes variation due to voters who may later change their minds.

When the voting population is large then the noise induced by ϵ^i in the mapping between the population parameters v_r and x_r is negligible,[30] and so it follows that:

$$x_r \approx \Phi\left(\frac{v_r - 0.5}{\sigma_\epsilon}\right) \quad [5]$$

From this, we can back out the implied expected vote share by inverting this function:

$$E[v_r|x_r] \approx 0.5 + \sigma_\epsilon \Phi^{-1}(x_r) \quad [6]$$

This function is also shown as a dashed line in Figure 2, based on our estimated value of $\widehat{\sigma_\epsilon} = 0.150$. To be clear, this is the appropriate mapping only if we know the true population proportion who expect a particular candidate to win, x_r. While this assumption is clearly false, our goal here is to provide a forecast comparable to the naïve forecast of voter intentions, and so in both cases we use the mapping between survey proportions and forecasts that would be appropriate in the absence of sampling variation. Indeed, in Figure 2 many of the extreme values of the proportion of voters expecting a candidate to win likely reflect sampling variation. That is, our transformation of voter expectations data into vote shares is clearly not estimated as a line of best fit, as elections are rarely as lopsided as the dashed line suggests. This feature roughly parallels the observation that in Figure 1; elections are rarely as lopsided as suggested by small samples of voter intentions. We will explore this feature of the data in greater detail in the next section when we evaluate efficient shrinkage-based estimators. But for now we will call this simple transformation of voter expectation our "naïve" expectation-based forecast.

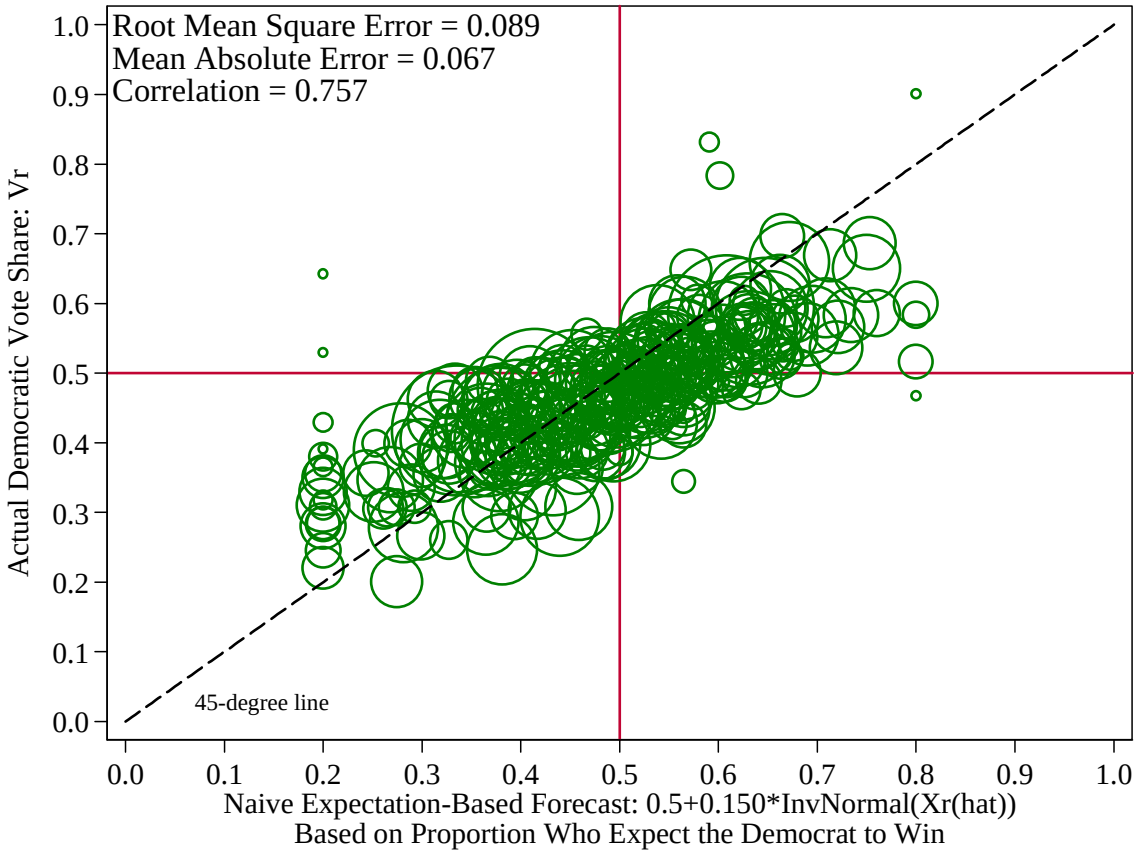
The one remaining difficulty is that in 22 races (7 percent of races), either 0% or 100% of survey respondents expect the Democrat to win, and so equation [6] does not yield a specific forecast. In these cases we (somewhat arbitrarily) infer that the

[30] As we will see in the next section, the noise term ϵ_r^i is relevant to the mapping between the population parameter x_r and the sample estimate $\widehat{x_r}$.

candidate is expected to win 20% or 80% of the vote, respectively.[31] Given the extreme nature of these inferences, we regard these assumptions as unfavorable to the expectations-based forecast. Even so, we obtain qualitatively similar results when imputing expected vote shares of 0% and 100% instead. (Section IV provides a more satisfactory treatment of this issue.)

Figure 3 plots our naïve expectations-based forecasts of vote shares against the actual election results. These forecasts are clustered along the 45-degree line, suggesting that they are quite accurate.

Figure 3: Naïve Expectation-Based Forecast and Actual Vote Share



[31] Our rationale was simply that these are the nearest round numbers that ensure that the implied forecast is monotonic in the proportion of respondents expecting a particular candidate to win. (Across the 289 other elections, the minimum was 24% and the maximum was 76%.)

In Table 3 we provide several simple comparisons of forecast accuracy. The first two rows show that the expectation-based forecast yields both a root mean squared error and mean absolute error that is significantly less than the intention-based forecast. The third row shows that the expectation-based forecast is also the more accurate forecast in 65% of these elections. The significance in these advantages for expectation-based forecasts is shown in the corresponding final column. In the fourth row, we examine the correlation coefficient. One might be concerned that the better performance of the expectation-based forecasts reflects the fact that they rely on an estimated parameter, σ_ϵ, and thus they use up one more degree of freedom. That is, our estimated value of σ_ϵ "tilts" the expectations data so that the implied forecasts lie along the 45-degree line. The correlation coefficient effectively both tilts the data and shifts it up and down, so as to maximize fit. Thus, it arguably puts each forecast on something closer to an equal footing. Even so, the expectation-based forecasts are also more highly correlated with actual vote shares than are the intention-based forecasts.

Table 3: Comparing the Accuracy of Naïve Forecasts of Vote Shares

	Raw Voter Intention: $\widehat{v_r}$	Transformed Voter Expectation: $\mathbf{0.5 + 0.150\Phi^{-1}(\widehat{x_r})}$	Test of Equality
Root Mean Squared Error	0.151 (0.008)	0.089 (0.005)	t_{344}=7.05 (p<0.0001)
Mean Absolute Error	0.115 (0.006)	0.067 (0.003)	t_{344}=8.81 (p<0.0001)
How often is forecast closer?	35.0% (2.7)	65.0% (2.7)	t_{344}=5.37 (p<0.0001)
Correlation	0.571	0.757	
Encompassing regression: $v_r = \alpha + \beta_v Intention_r + \beta_x Expectation_r$	0.058** (0.026)	0.480*** (0.035)	
Optimal weights: $v_r = \beta Intention_r + (1-\beta) Expectation_r$	8.5%** (3.7)	91.5%*** (3.7)	

Notes: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10%, respectively. (Standard errors in parentheses). These are assessments of forecasts of the Democrat's share of the two-party vote in n=311 elections. Comparisons in the third column test the equality of the measures in the first two columns. In the encompassing regression, the constant $\hat{\alpha} = 0.207^{***}$ (se=0.013).

We also show a Fair-Shiller (1989 and 1990) regression, attempting to predict election outcomes on the basis of a constant, and our two alternative forecasts. The expectation-based forecast has a large and extremely statistically significant weight, as does the constant. But it does not fully encompass the information in the intention-based forecast, which still receives a statistically significant, albeit small weight. Finally, we estimate the optimal weighted average of these two forecasts, which puts greater than 90% of the weight on the expectation-based forecast.

These tests, of course, are all based on the common but problematic assumption that our sample data can be interpreted as representing population moments. We now

turn to generating statistically efficient forecasts instead, and will repeat our forecast evaluation exercise based on these adjusted forecasts.

IV. Efficient Poll-Based Forecasts of Vote Share

An important insight common to both Figure 1 and Figure 3 is that in those elections where the Democrats are favored, the final outcome typically does favor Democrats, but by less than suggested by the naïve forecasts based on either voter intentions or voter expectations. Likewise, when the poll favors Republicans, the Democrats do tend to lose, but again, by less than suggested by our naïve forecasts. In fact, this is a natural implication of sampling error, because our extreme polls likely reflect sampling variation, and so it is unsurprising that they are not matched by extreme outcomes. It is this observation that motivates our use of shrinkage estimators in this section—"shrinking" the raw estimates of the proportion intending to vote for one candidate or another toward a closer race. This idea is widely understood by political scientists (Campbell, 2000), but is typically ignored in media commentary. We will also adjust for any biases (or "house effects") in these data.

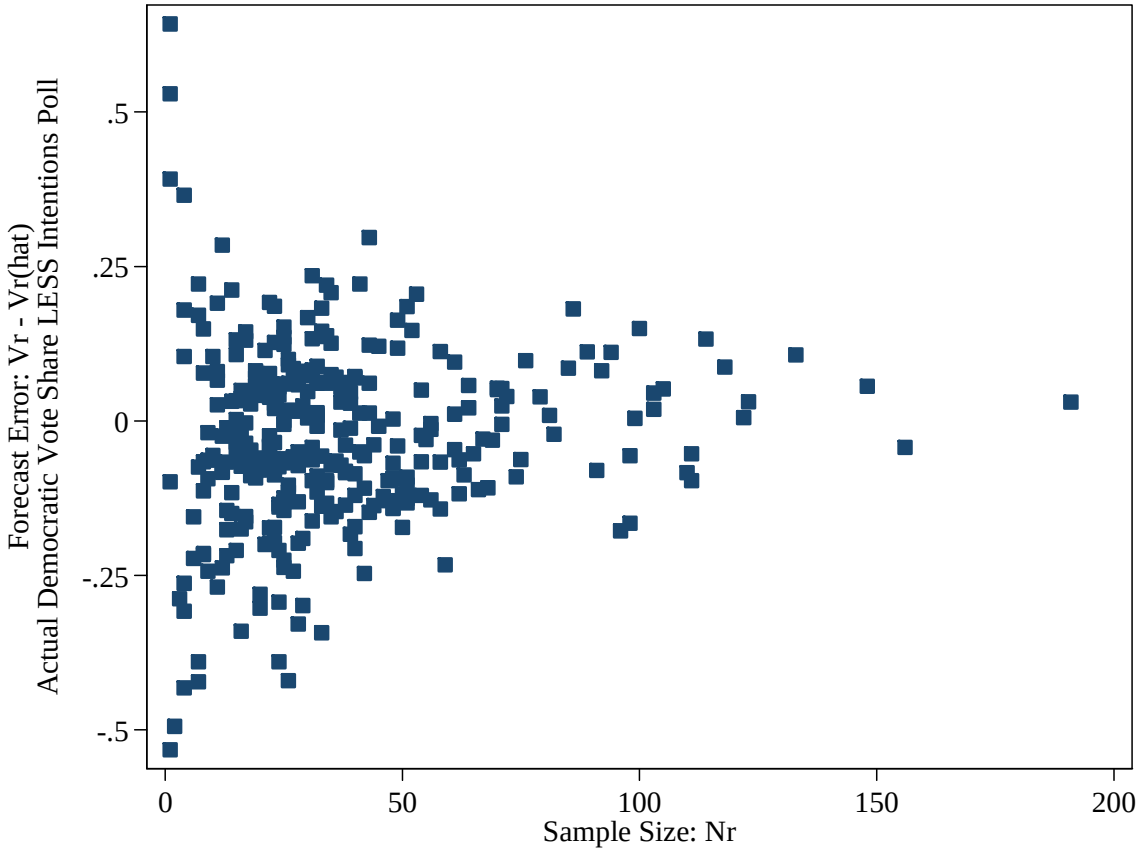
In the following discussion it is important to distinguish between the actual vote share won by the Democrat, v_r, from the sample proportion who intend to vote for the candidate, $\widehat{v_r}$, and the optimal intention-based forecast, $E[v_r|\widehat{v_r}]$. Likewise, we distinguish the sample proportion who expect a candidate to win, $\widehat{x_r}$, from the population proportion, x_r, and our optimal expectation-based forecast of the vote share, $E[v_r|\widehat{x_r}]$. Because we are only analyzing respondents with valid expectation and intent data, each forecast will be based on the same sample size, n_r.

We will begin by analyzing forecasts based on standard polls of voter intentions, and will then turn to analyzing how voter expectations might improve these forecasts.

Interpreting Voter Intentions

Our goal is to find the mapping between our raw voter intentions data, and the forecast which minimizes the mean squared forecast error. The usual approach—of fitting an OLS regression line—involves shrinking each estimate back toward the grand mean, using the signal-to-noise ratio. (This is why the least-squares estimator of an errors-in-variables model yields a regression coefficient that is shrunk by a factor related to the signal-to-noise ratio in the explanatory variable.) The difficulty is that in our setting, the sample size varies widely across each race, and as Figure 4 illustrates, so too does the noise underpinning each observation.

Figure 4: Sample Size and Forecast Errors in the Intent Poll



Our point of departure in this section is to note that our sample estimate of the proportion of voters intending to vote for a candidate is a noisy estimate—and possibly also biased—estimate of the election outcome:

$$\widehat{v_r} = v_r + b + e_r, \text{ where } e_r \sim N(0, \sigma_{e_r}^2) \quad [7]$$

Where and b is a bias term which picks up the pro-Democrat house effect in ANES polls, and e_r is the noise term.[32] In particular, notice the r subscript on the variance of this noise term, which reflects the fact that sampling variability will vary with the characteristics (particularly, sample size) of a specific poll. Assuming that $E[v_r e_r] = 0$ we get the following familiar result:

$$E[v_r|\widehat{v_r}] = \mu_v + \frac{\sigma_v^2}{\sigma_{\widehat{v}}^2}(\widehat{v_r} - b - \mu_v) \quad [8]$$

where μ_v and σ_v^2 are the mean and variance of the Democratic vote share, across all the races in our dataset. Forming an optimal intentions-based forecast requires estimating each of these parameters. We estimate the average vote share of Democrats directly from our sample: $\widehat{\mu_v} = \frac{1}{R}\sum v_r = 0.468$ $(se = 0.005)$, and the variance of the Democrat vote share is: $\widehat{\sigma_v^2} = \frac{1}{R-1}\sum(v_r - \widehat{\mu_v})^2 = 0.0089$. Likewise, it is easy to estimate the bias term, $\hat{b} = \sum(\widehat{v_r} - v_r)/R = 0.031$ $(se = 0.008)$. (This bias in the ANES is reasonably large, statistically significant, and to our reading, has not previously been documented; even when we cluster results by year, the bias remains statistically significant.) All that remains is to sort out the variance of the polls, which can be broken into: $\sigma_{\hat{v}}^2 = \sigma_v^2 + \sigma_e^2$, where $\sigma_e^2 = E[e_r^2] = E[(v_r - \widehat{v_r} - b)^2]$.[33]

[32] We also tested for an anti-incumbent party bias, but found it to be small and statistically insignificant.

[33] Any variance in the bias term from cycle to cycle would be included in the variance of the polling error.

There are two sources of error to consider. First, these polls are typically taken one month prior to the election,[34] and many voters may change their minds in the final weeks of the campaign. Hence while we are sampling from a population where v_r percent of respondents will ultimately vote for the Democrat, but when they are polled one month prior, an extra τ_r percent intend to vote Democratic. Second, sampling error plays an important role, particularly in small samples. Assuming that these two sources of error are orthogonal so that $E[\tau_r(v_r - \widehat{v_r} - b - \tau_r)] = 0$, the variance of the polling error can be decomposed as:

$$\sigma_e^2 = E[(v_r - \widehat{v_r} - b)^2] = E[(v_r - (v_r + \tau_r))^2] + E\left[\left((v_r + \tau_r) - (\widehat{v_r} - b)\right)^2\right] \quad [9]$$

where the first term in the above expression reflects the variability in "true" public opinion between polling day and Election Day (and hence is unrelated to the size of our samples). Our data does not have any useful time variation, and so we simply use the estimate of $\widehat{\sigma_\tau^2} = 0.001$, from Lock and Gelman (2010) (who estimate this as a function of time before the election; we use their fitted value for one-month before Election Day). The second term reflects sampling variability, and because the poll result $\widehat{v_r}$ is the mean of a binomial variable with mean $v_r + \tau_r$, this second term can be expressed as:

$$E[(v_r - \tau_r - b - \widehat{v_r})^2] = \frac{(v_r + \tau_r)(1 - v_r - \tau_r)}{n_r^{eff}} \approx \frac{0.25}{n_r / \gamma_r^v} \quad [10]$$

The numerator of this expression is the product of the vote shares of the two parties, if the election were held on polling day. For most elections (and particularly competitive contests), the term $(v_r + \tau_r)(1 - v_r - \tau_r) \approx 0.25$, and so we use this approximation. (We obtain similar results when we plug in actual vote shares or vote shares from the previous election instead; our approximation has the virtue of being usable in a real-time forecasting context.) The denominator, n_r^{eff} denotes the effective sample size of the specific poll in race r. If the sample were a simple random sample,

[34] In fact, the polls are taken fairly uniformly over the two months prior to the election.

the effective sample size would be exactly equal to the actual sample size. But the American National Election Studies uses a complex sample design, polling only in a limited number of primary sampling units. The "design effect" γ_r^v corrects for the effects of the intra-cluster correlation within these sampling units. (The subscript r serves as a reminder that it varies with the sample size of specific poll—for instance, it is one when $n_r = 1$—and the superscript v serves as a reminder that the design effect varies, and this is the design effect for voter intentions). Unfortunately published estimates of γ_r^v are based on design effects in national samples, and so they cannot be applied to our analysis of state samples. Moreover, the public release files in the ANES files do not contain sufficient detail about the sampling scheme to allow us to estimate these design effects directly. A standard approach to estimating γ_r^v is by the so-called "Moulton factor", $\gamma_r^v = 1 + (n_r - 1)\rho$, where ρ is the intra-cluster correlation coefficient (Moulton, 1990).[35] In what follows, we assume that ρ is constant across states and time. While we lack the details on sampling clusters to estimate ρ directly, we can estimate it indirectly. Figure 4 highlights the underlying variation, showing a consistent pattern of errors varying with the sample size of a poll. Our identification comes from the fact that this pattern is shaped by ρ—the higher is ρ, the less quickly the variance of $(v_r - \widehat{v_r})$ declines with sample size. Thus, we return to equation [8], plug in the values for $\widehat{\mu_v}$, $\widehat{\sigma_v^2}$, $\widehat{\sigma_t^2}$ and $\hat{b}$ and estimate ρ directly by running non-linear least-squares on the following regression:

$$v_r = \widehat{\mu_v} + \frac{\widehat{\sigma_v^2}}{\widehat{\sigma_v^2} + \widehat{\sigma_t^2} + \frac{1 + (n_r - 1)\rho}{4n_r}} (\widehat{v_r} - \hat{b} - \widehat{\mu_v}) \qquad [11]$$

which yielded an estimate of $\hat{\rho} = 0.030$ (with a standard error of 0.0080), which implies an average design effect of $\widehat{\gamma^v} = 2.09$. Thus, returning to equation [8], our MSE-minimizing forecast based on the voter intentions data, is:

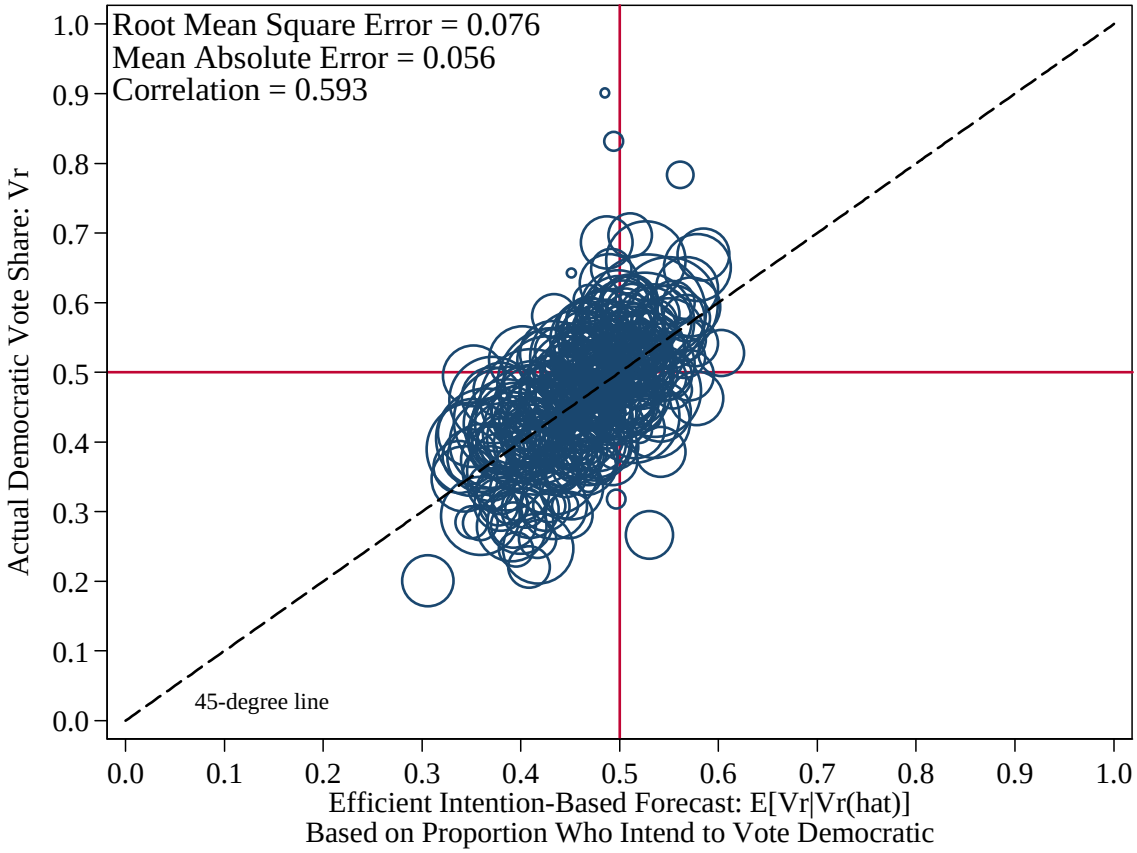
[35] A remaining difficulty is that while we assume that the total observations for each state come from a single cluster, in fact some of the larger states include multiple primary sampling units. Thus, our approach is appropriate if we think of ρ as the intra-cluster correlation, where a "cluster" is the set of sampled addresses within a state.

$$E[v_r|\widehat{v_r}] = \mu_v + \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2}\,(\widehat{v_r} - b - \mu_v)$$

$$= 0.468 + \frac{0.0089}{0.0089 + 0.0001 + \frac{1 + (n_r - 1)0.030}{4n_r}}(\widehat{v_r} - 0.031 - 0.468)$$

Given the data in our sample, the shrinkage estimator (the coefficient on the de-meaned and de-biased intent poll) averages 0.33 (which corresponds closely with the average slope seen in Figure1, but it ranges from 0.03 (in a race with only one survey respondent) to 0.47.

Figure 5: Efficient Intention-Based Forecasts and Actual Vote Share



In Figure 5, we show the relationship between our optimal intention-based forecast and actual vote share. These adjusted intention-based forecasts are clearly more accurate than the naïve forecasts numbers: the forecasts lie along the 45-degree line, and

both the mean absolute error and the root mean squared error are about half that found Figure 1. We now turn to finding the most efficient transformation of our voter expectation data.

Interpreting Voter Expectations

In our previous analysis in Section III, we transformed data on voter expectations into vote share forecasts based on $E[v_r|x_r]$. But taking the sample variability seriously means that we are trying to figure out $E[v_r|\widehat{x_r}]$. As an intermediate step, we begin by estimating the $E[x_r|\widehat{x_r}]$. Once again, we turn to a shrinkage estimator in order to generate efficient estimates of x_r, given our small sample estimates, $\widehat{x_r}$. As in equation [8],

$$E[x_r|\widehat{x_r}] = \mu_x + \frac{\sigma_x^2}{\sigma_{\hat{x}}^2}\,(\widehat{x_r} - b - \mu_x) \qquad [12]$$

where μ_x is the mean across all elections of the proportion of the population who expect the Democrat to win; σ_x^2 of the corresponding variance, while $\sigma_{\hat{x}}^2$ is the variance of the corresponding sample estimator; and as before, we have a bias parameter, b which allows for the possibility that these data oversample people who expect Democrats to win.

The key difficulty in working with data on voter expectations rather than voter intentions is that while we do ultimately observe how the entire population votes, we never observe what the whole population expects. Thus in estimating population parameters, we will rely heavily on the mapping in equation [5] between population vote shares and population expectations.[36] Using this insight, we estimate $\widehat{\mu_x} = \sum \Phi\left(\frac{v_r - 0.5}{\widehat{\sigma_\epsilon}}\right)/R = 0.427\ (se = 0.005)$. Likewise we estimate the bias term $\hat{b} = \sum (\widehat{x_r} - \Phi\left(\frac{v_r - 0.5}{\widehat{\sigma_\epsilon}}\right))/R = 0.043\ (se = 0.009)$. This bias term represents the increased

[36] We are yet to adjust our standard errors for the extra uncertainty generated because we are using an estimate of $\widehat{\sigma_\epsilon}$.

probability of an individual expects a Democrat to win, its importance cannot be directly compared to the intention data, where the bias has a meaningful impact in the naïve intention-based forecast of vote share.

The numerator of the shrinkage estimator in equation [12] is the variance of the population expectation across all elections, and it is also quite straightforward to estimate: $\widehat{\sigma_x^2} = \frac{1}{R-1}\sum(x_r - \widehat{\mu_x})^2 = 0.0450$. All that remains is to sort out the denominator of the shrinkage estimator, $\sigma_{\hat{x}}^2$, which is equal to the underlying variation in population expectations across elections, plus sampling variability. Because different elections have different sample sizes, this denominator varies across elections. Again, because we are asking about binary outcome—whether or not you expect the Democrat to win in your state—we know the functional form of the relevant sampling error. Thus:

$$\sigma_{\hat{x}}^2 = \sigma_x^2 + \frac{x_r(1-x_r)}{n_r^{eff}} \approx \frac{0.25}{n_r/\gamma_r^x} \qquad [13]$$

where, the approximation follows because the product of the population proportions expecting each candidate, $x_r(1-x_r) \approx \Phi\left(\frac{v_r-0.5}{\widehat{\sigma_\epsilon}}\right)\Phi\left(\frac{0.5-v_r}{\widehat{\sigma_\epsilon}}\right)$, and in turn, $\Phi\left(\frac{v_r-0.5}{\widehat{\sigma_\epsilon}}\right)\Phi\left(\frac{0.5-v_r}{\widehat{\sigma_\epsilon}}\right) \approx 0.25$ because most elections are competitive (and $\widehat{\sigma_\epsilon}$ is not too small).[37] As before, n_r^{eff} is the effective sample size, and γ_r^x is the relevant design effect, and we apply the "Moulton factor" to estimate $\gamma_r^x = 1 + (n_r - 1)\rho_x$. The x superscript on the intra-cluster correlation reminds us that there is no reason to expect the intra-cluster correlation in voter expectations to be similar to that for voter intentions. Thus, our remaining challenge is to estimate the intra-class correlation in voter expectations. As we did with the voter intentions data, we will use an indirect approach to estimating

[37] Notice that equation [13] involves the product of the *population* proportions who expect each candidate to win, and it is this product that ≈0.25. As previously noted, we don't actually observe these population proportions, but we do observe the population vote shares, and using the transformation in equation [5], we confirm that $\Phi\left(\frac{v_r-0.5}{\widehat{\sigma_\epsilon}}\right)\left(1-\Phi\left(\frac{0.5-v_r}{\widehat{\sigma_\epsilon}}\right)\right) \approx 0.25$. In our small samples, $\widehat{x_r}(1-\widehat{x_r})$ can diverge quite significantly from 0.25.

ρ_x. That is, we plug the results in equations [12] and [13] back in to equation [6] to get a simple vote share forecasting equation, and find the value that yields the best overall fit:

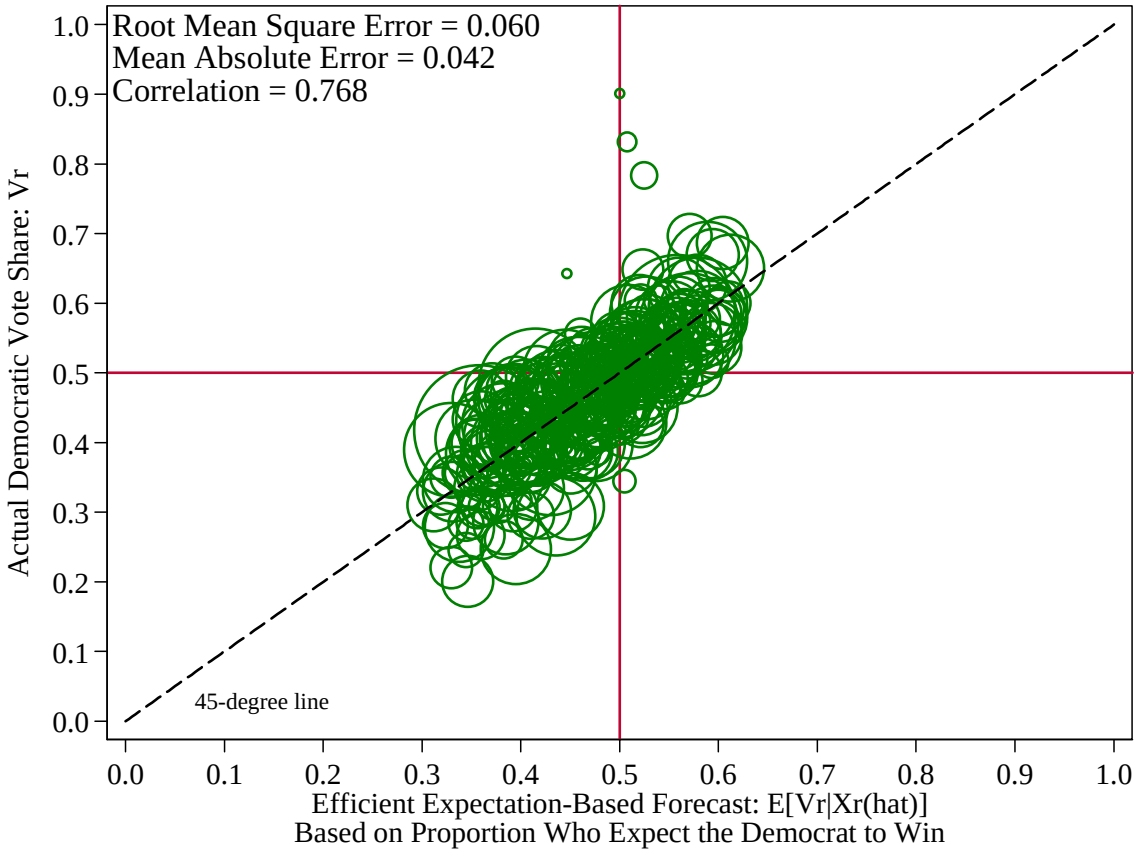
$$E[v_r|\widehat{x_r}] = 0.5 + \widehat{\sigma_\epsilon}\Phi^{-1}\left(\widehat{\mu_x} + \frac{\widehat{\sigma_x^2}}{\widehat{\sigma_x^2} + \frac{1+(n_r-1)\rho_x}{4n_r}}\left(\widehat{x_r} - \hat{b} - \widehat{\mu_x}\right)\right) \quad [14]$$

We fit this equation using non-linear least squares, and it yields an estimate of $\widehat{\rho_x} = 0.045$ (se $= 0.010$), which in turn implies an average design effect $\widehat{\gamma^x} = 2.62$, and that the shrinkage estimator which has an average value of 0.65, ranges from 0.14 to 0.76. Thus, our MSE-minimizing forecast of the Democrats vote share, based only the voter expectations data is:

$$E[v_r|\widehat{x_r}] = 0.5 + 0.150\Phi^{-1}\left(0.440 + \frac{0.040}{0.040 + \frac{1+(n_r-1)0.045}{4n_r}}(\widehat{x_r} - 0.043 - 0.427)\right)$$

In Figure 6, we show the relationship between our efficient expectation-based forecast and actual election outcomes. These adjusted expectation-based forecasts are clearly very accurate, and appropriately scaled: the forecasts lie along the 45-degree line.

Figure 6: Efficient Expectation-Based Forecast and Election Outcomes



Forecast evaluation

Turning to Table 4, we see that these efficient forecasts based on voter expectations are more accurate than efficient intention-based forecasts. Again, the first two rows show that the expectation-based forecasts yield both a root mean squared error and mean absolute error that is less than the intention-based forecasts by a statically significant amount. The third row shows that the expectation-based forecasts are also the more accurate forecast in 63% of these elections, very similar to the naïve approach. The expectation-based forecasts are still more highly correlated with actual vote shares than are the intention-based forecasts by a sizable margin. In the Fair-Shiller regression, the expectation-based forecasts have a much larger weight than the intention-based forecast; both are statistically significant, so the intention-based data is providing some unique information. Finally, an optimally weighted average still puts

just over 90% of the weight on the expectation-based forecast. While that number may not seem remarkable in the context of this paper, it is remarkable in the context of society. If you take step back from this paper, consider that the average weight placed on expectation polls by pollsters, the press, and campaigns is 0%, as it is generally ignored.

Table 4: Comparing the Accuracy of Efficient Forecasts of Vote Share

	Efficient Voter Intention: $E[v_r \mid \widehat{v_r}]$	Efficient Voter Expectation: $E[v_r \mid \widehat{x_r}]$	Test of Equality
Root Mean Squared Error	0.076 (0.005)	0.060 (0.006)	t_{310}=5.75 (p<0.0001)
Mean Absolute Error	0.056 (0.003)	0.042 (0.002)	t_{310}=6.09 (p<0.0001)
How often is forecast closer?	37.0% (2.6)	63.0% (2.6)	t_{310}=4.75 (p<0.0001)
Correlation	0.593	0.768	
Encompassing regression: $v_r = \alpha + \beta_v Intention_r + \beta_x Expectation_r$	0.184** (0.089)	0.913*** (0.067)	
Optimal weights: $v_r = \beta Intention_r + (1-\beta) Expectation_r$	9.5% (6.7)	90.5%*** (6.7)	

Notes: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10%, respectively. (Standard errors in parentheses). These are assessments of forecasts of the Democrat's share of the two-party vote in n=311 elections. Comparisons in the third column test the equality of the measures in the first two columns. In the encompassing regression, the constant $\hat{\alpha} = -0.046$ (se=0.030).

V. Efficient Poll-Based Probabilistic Forecasts

So far we have introduced two different outcome variables: the prediction of the winner and the level of the outcome, but frequently the most desirable outcome metric for stakeholders of an event is the probabilistic forecast. In this section we translate our raw polling data, from both voter intention and expectation, into efficient poll-based

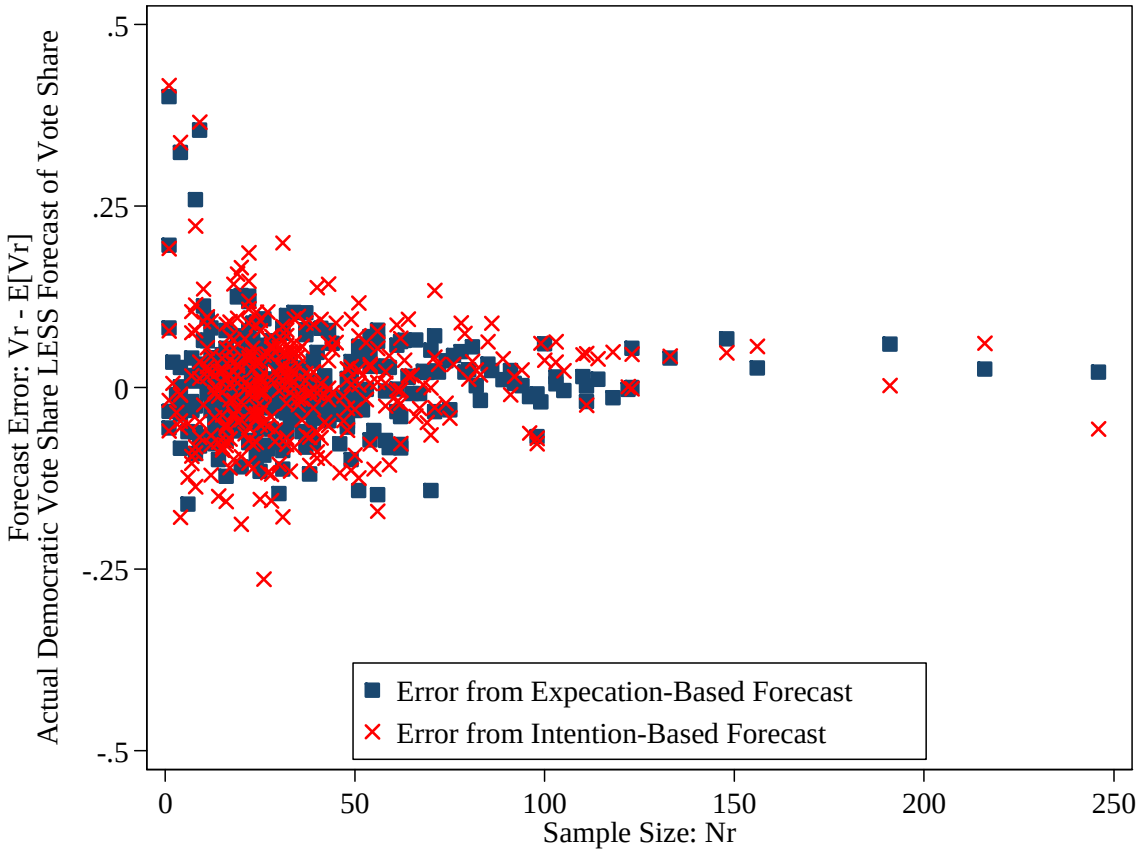
probabilistic forecasts. We start with our forecast of the vote shares generated from both types of data: $E[v_r|\widehat{v_r}]$ for intention data and $E[v_r|\widehat{x_r}]$ for expectation data. The probability that the Democrat wins the race is the probability that the actual vote share is greater than 50%.[38] We assume that the errors of our expectations are normally distributed and the probability of victory for the Democratic candidate becomes:

$$Prob(v_r > 0.5) = \Phi\left(\frac{E[v_r] - 0.5}{\sigma_{\varepsilon,E[v_r]}}\right) \quad [15]$$

To create probabilistic forecasts, we are going to need to determine the variance of the error. Figure 7 revisits Figure 4, but instead of showing the errors from the raw intention poll, it shows the errors from the efficient intention and expectation-based forecasts. The figure illustrates that there is still noise and that it varies with sample size.

[38] This is not true in many types of elections, but it is true in almost all Electoral College races in our dataset.

Figure 7: Sample Size and Forecast Errors in the Forecasts of Vote Share



To determine the variance of the error, we start with the standard error of a forecast: $se = \frac{RMSE}{\sqrt{n}}\left(1 + \frac{(X_r - \mu_x)^2}{\sigma_x^2}\right)^{1/2}$ (Barreto and Howland, 2006). We use equation [11], $v_r = \widehat{\mu_v} + \frac{\widehat{\sigma_v^2}}{\widehat{\sigma_v^2} + \widehat{\sigma_\tau^2} + \frac{1+(n_r-1)\rho}{4n_r}}(\hat{v}_r - \hat{b} - \widehat{\mu_v})$ and we think of the $(\hat{v}_r - \hat{b} - \widehat{\mu_v})$ as the raw data X_r, where μ_x is the average raw data. Thus, the variance of the forecasts:

$$\widehat{\sigma^2_{\varepsilon,E[v_r]}} = RMSE^2 + \frac{RMSE^2}{\sigma^2_{poll}}\left(poll - \mu_{poll}\right)^2 = RMSE^2 + \sigma^2_{Shrinkage}\left(poll - \mu_{poll}\right)^2 \quad [16]$$

The root mean square error (squared) is of the two equations we use to determine the shrinkage estimate; equation [11] for intention and equation [14] for expectation. It captures the accuracy of the equation in estimating the within sample data used to calibrate the forecast model. The other part of the variance is the standard deviation of the shrinkage estimator (squared) interacted with the distance between the raw data and the mean data. The intuition behind this is that we should assume less

accurate forecasts, the less accurate the estimation of the coefficient we use to transform our raw data is and how unusual the raw data is compared to the average raw data. For the expectation data, we use the "naïve" forecast of vote share as the raw data. This term is going to pick up the noise correlated with sample size, as low sample sizes have more polls that post very far from the average poll.

The RMSE for intention is 0.076 and expectation 0.060; as we know from Section IV, the forecasts of vote share from the expectation data is more accurate. $\frac{RMSE^2}{\sigma^2_{poll}} = \sigma^2_{Shrinkage}$ is 0.177 for intention and 0.186 for expectation. On average, about half of the $\sigma_{\varepsilon,E[v_r]}$ comes from the each side of the equation and the estimated standard deviation does decrease with sample size. Putting together equation [15] and [16]:

$$Prob(v_r > 0.5) = \Phi\left(\frac{E[v_r] - 0.5}{\sqrt{RMSE^2 + \frac{RMSE^2}{\sigma^2_{poll}}\left(poll - \mu_{poll}\right)^2}}\right) \quad [17]$$

In Table 5 we show that probabilistic forecasts based on the expectation data are more accurate than those based on the intention data. The first row shows that the expectation-based probabilities yield a smaller root mean squared error than the intention-based probabilities by a statically significant amount. The second row shows that the expectation-based probabilities are also the more accurate forecast in 80% of these elections, a statistically significant difference. In the Fair-Shiller regression, the expectation-based probabilities have a much larger weight than the intention-based forecast; the intention-based data is not statistically significant, so we cannot discount the possibility that the intention-based probability is providing no unique information. Finally, an optimally weighted average puts most of the weight on the expectation-based probability.

Table 5: Comparing the Accuracy of Efficient Probabilistic Forecasts

	Efficient Voter Intention: $\boldsymbol{Prob(v_r > 0.5 \mid \widehat{v_r})}$	Efficient Voter Expectation: $\boldsymbol{Prob(v_r > 0.5 \mid \widehat{x_r})}$	Test of Equality
Root Mean Squared Error	0.442 (0.007)	0.345 (0.012)	t_{344}=10.4 (p<0.0001)
How often is forecast closer?	20.0% (2.3)	80.0% (2.3)	t_{344}=13.2 (p<0.0001)
Encompassing regression: $\boldsymbol{I(DemWin)_r = \Phi\left(\alpha + \beta_v \Phi^{-1}(Prob_I) + \beta_x \Phi^{-1}(Prob_x)\right)}$	0.216 (0.155)	1.776*** (0.201)	
Optimal weights: $\boldsymbol{I(DemWin)_r = \Phi\left(\beta \Phi^{-1}(Prob_I) + (1-\beta)\Phi^{-1}(Prob_x)\right)}$	-0.784*** (0.185)	1.784*** (0.185)	

Notes: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10%, respectively. (Standard errors in parentheses). These are assessments of forecasts of the Democrat's probability of victory in n=311 elections. Comparisons in the third column test the equality of the measures in the first two columns. In the encompassing regression, the constant $\hat{\alpha} = -0.035$ (se=0.114).

VI. Efficient Poll-Based Forecasts of 2008

We now test how the coefficients, developed with data from 1952-2004, forecast the 2008 Electoral College races. Since this is an "out-of-sample" comparison, the forecast of vote share use the coefficients derived in Section IV and the probabilistic forecast use the coefficients derived in Section V. In Table 2 we show that 2008 is just one of two years in which the expectation poll is not more correct in predicting the winner of the races than the intention poll; they were tied in 2008. Yet, Table 6 shows that the expectation question provides a much more accurate and informative forecast of 2008 than the intention question. The expectation-based forecasts have smaller errors and higher correlation in all of the forecast for vote share and probability of victory comparisons, and carry more weight in all of the joint estimations for forecast of vote share and partially in probability of victory. Since the expectation-based forecasts are

statistically significant in encompassing regression and the intention-based forecast is not, we cannot rule out that intention data provides no useful information beyond expectation data in 2008.

Table 6: Comparing the Accuracy of Efficient Forecasts for the 2008 Electoral College

Forecast of Vote Share:	Efficient Voter Intention: $E[v_r \mid \widehat{v_r}]$	Efficient Voter Expectation: $E[v_r \mid \widehat{x_r}]$	Test of Equality
Root Mean Squared Error	0.093 (0.021)	0.085 (0.022)	t_{33}=1.28 (p<0.2105)
Mean Absolute Error	0.063 (0.012)	0.056 (0.011)	t_{33}=0.92 (p<0.3656)
How often is forecast closer?	47.1% (8.7)	52.9% (8.7)	t_{33}=0.34 (p<0.7371)
Correlation	61.6%	69.2%	
Encompassing regression: $v_r = \alpha + \beta_v Intention_r + \beta_x Expectation_r$	0.330 (0.291)	0.684*** (0.250)	
Optimal weights: $v_r = \beta Intention_r + (1-\beta) Expectation_r$	24.7% (26.7)	75.3%*** (26.7)	
Probabilistic Forecasts:	$Prob(v_r > 0.5 \mid \widehat{v_r})$	$Prob(v_r > 0.5 \mid \widehat{x_r})$	
Root Mean Squared Error	0.458 (0.022)	0.403 (0.048)	t_{344}=1.55 (p<0.1295)
How often is forecast closer?	23.5% (7.4)	76.5% (7.4)	t_{344}=3.58 (p<0.0011)
Encompassing regression: $I(DemWin)_r = \Phi\left(\alpha + \beta_v \Phi^{-1}(Prob_I) + \beta_x \Phi^{-1}(Prob_x)\right)$	1.618 (1.289)	1.224** (0.520)	
Optimal weights: $I(DemWin)_r = \Phi\left(\beta \Phi^{-1}(Prob_I) + (1-\beta)\Phi^{-1}(Prob_x)\right)$	2.4% (39.1)	97.6%** (39.1)	

Notes: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10%, respectively. (Standard errors in parentheses). These are assessments of forecasts of the

Democrat's forecast of voter share and probability of victory in n=34 elections (the 34 Electoral College races chosen by the ANES). Comparisons in the third column test the equality of the measures in the first two columns. In the encompassing regression, the constant $\hat{\alpha} = 0.032$ (se=0.097) for vote share and $\hat{\alpha} = 1.242$ (se=0.504) for probability of victory.

VII. Efficient Poll-Based Forecasts from Non-Random Samples

Under many circumstances we are left making forecasts based off polls with overtly non-random samples. An extreme example of that would be a poll where all of the respondents support just one of the candidates. In this section we test the accuracy of forecasts based off of the expectation question data originating from respondents who are exclusively voting for one major party or the other, by comparing those forecasts to forecasts created from the full sample of intentions. For the non-random samples, the sample size drops from an average of 38 to about 19 respondents.

We start by revisiting equation [12] from Section IV: $E[x_r|\widehat{x_r}] = \mu_x + \frac{\sigma_x^2}{\sigma_{\hat{x}}^2}(\widehat{x_r} - b - \mu_x)$, where μ_x is the mean across all elections of the proportion of the population who expect the Democrat to win and σ_x^2 is the corresponding variance. These values are the same in this biased sample as they are in Section IV.[39] $\sigma_{\hat{x}}^2$ is the variance of the sample estimator, and our bias parameter, b, will attempt to account for the oversample people who expect Democrats to win (which should be heavily positive in a Democratic only sample and heavily negative in a Republican only sample).

The bias term is $\hat{b} = \sum(\widehat{x_r} - \Phi\left(\frac{v_r - 0.5}{\widehat{\sigma_\epsilon}}\right))/R$. This is 0.203 (0.016) for the Democratic sample and -0.112 (0.012) for the Republican sample. In Section IV, we conclude that the full dataset has a bias of 0.042, approximately half the difference in the absolute bias of the Democratic and Republican supporters.

[39] There is a slight variation in the variables because 5 of the races in our dataset have no Democratic supporters and 4 no Republican supporters. Thus, when we drop those races from the dataset when we explore forecast accuracy of the respective non-random samples.

We make the same assumptions as in Section IV in regard to $\widehat{\sigma_{\hat{x}}^2}$, which, combining equations [13] and [14], is $\widehat{\sigma_x^2} + \frac{(1+(n_r-1)\rho_x)}{4n_r}$. We are going to have to re-derive ρ for these two biased samples, as the clustering is going to have a totally different affect within party. Both samples have the same $\widehat{\sigma_x^2}$ and about half the observations as the full samples. This yields shrinkage estimators (i.e., $\frac{\sigma_x^2}{\sigma_{\hat{x}}^2}$) for the Democratic sample with an average of 0.59, ranges from 0.14 to 0.74. Not surprisingly, the full dataset has shrinkage estimators that are on average 7.5 percentage points larger, ranging from a low of the same to a higher of 54 percentage points larger. For the Republican sample, the differences are not as stark, with the Republican sample having slightly smaller $\widehat{\rho_x}$ than the full sample; the full sample has shrinkage estimators that are on average 2.5 percentage points larger, but they range from a low of -7 percentage points smaller to 42 percentage points larger.

In Table 7 we show that forecasts based on the expectations of the Democratic voters only or the Republican voters only, a non-random selection of approximately half of the sample, yield a lower root mean squared error, mean absolute error, and higher correlation than forecasts based on the full sample of voter intention. The first two rows show that the expectation-based forecasts yield both a root mean squared error and mean absolute error that is less than the intention-based forecasts by a statically significant amount for the Republican sample and a weakly significant amount for the Democratic sample. The third row shows that the expectation-based forecasts are also the more accurate forecast in the majority of elections. The expectation-based forecasts are still more highly correlated with actual vote shares than are the intention-based forecasts by a sizable margin. In the Fair-Shiller regression, the expectation-based forecasts have a much larger weight than the intention-based forecast; both are statistically significant, so the intention-based data may be providing some unique information. Finally, an optimally weighted average still puts nearly 60% of the weight

on the expectation-based forecast in the Democratic sample and just over 70% in the Republican sample.

Table 7: Comparing the Accuracy of Efficient Forecasts of Vote Shares from Biased Samples

Forecast of Vote Share:	Democratic Sample		Republican Sample	
	Intention: $E[v_r \mid \widehat{v_r}]$ Full Sample	Expectation: $E[v_r \mid \widehat{x_r}]$ Democratic Supporters	Intention: $E[v_r \mid \widehat{v_r}]$ Full Sample	Expectation: $E[v_r \mid \widehat{x_r}]$ Republican Supporters
Root Mean Squared Error	0.075 (0.005)	0.070 (0.006)	0.071 (0.004)	0.062 (0.004)
Mean Absolute Error	0.056 (0.003)	0.050 (0.003)	0.054 (0.003)	0.048 (0.002)
How often is forecast closer?	46.7% (2.9)	53.3% (2.9)	44.0% (2.8)	56.0% (2.8)
Correlation	0.592	0.664	0.604	0.718
Encompassing regression: $v_r = \alpha + \beta_v Intention_r + \beta_x Expectation_r$	0.625*** (0.078)	0.790*** (0.071)	0.489*** (0.077)	0.786*** (0.065)
Optimal weights: $v_r = \beta Intention_r + (1-\beta) Expectation_r$	38.5%*** (6.6)	61.5%*** (6.6)	29.8%*** (6.3)	70.2%*** (6.3)
	306 Elections		307 Elections	

Notes: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10%, respectively. (Standard errors in parentheses). These are assessments of forecasts of the Democrat's share of the two-party vote. In the encompassing regression with the Democratic sample, the constant $\hat{\alpha} = -0.196$ (se=0.036)*** and in the Republican sample, the constant $\hat{\alpha} = -0.129$ (se=0.031)***. A few of the 311 observations are dropped in each sample, because there is no intention for the given party.

In Table 8 we show that probabilistic forecasts based on the expectation data are more accurate than those based on the intention data. The first row shows that the expectation-based probabilities yield a smaller root mean squared error than the intention-based probabilities by a statically significant amount. The second row shows that the expectation-based probabilities are also the more accurate in slightly more

elections. In the Fair-Shiller regression, the expectation-based probabilities have a much larger weight than the intention-based forecast; both are statistically significant, so the intention-based data may be providing some unique information. Finally, an optimally weighted average still puts most of the weight on the expectation-based forecast.

Table 8: Comparing the Accuracy of Efficient Probabilistic Forecasts

Probabilistic Forecasts:	Democratic Sample: Voter Intention:	Democratic Sample: Voter Expectation:	Republican Sample: Voter Intention:	Republican Sample: Voter Expectation:
Root Mean Squared Error	0.444 (0.006)	0.388 (0.010)	0.442 (0.006)	0.357 (0.013)
How often is forecast closer?	28.4% (2.6)	71.5% (2.6)	19.9% (2.3)	80.1% (2.3)
Encompassing regression: $\boldsymbol{I(DemWin)_r = \Phi\left(\alpha + \beta_v \Phi^{-1}(Prob_I) + \beta_x \Phi^{-1}(Prob_x)\right)}$	1.73*** (0.40)	1.62*** (0.20)	1.29*** (0.41)	1.53*** (0.17)
Optimal weights: $\boldsymbol{I(DemWin)_r = \Phi\left(\beta \Phi^{-1}(Prob_I) + (1-\beta)\Phi^{-1}(Prob_x)\right)}$	-0.336** (0.170)	1.336*** (0.170)	-0.435*** (0.152)	1.435*** (0.152)
	306 Elections		307 Elections	

Notes: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10%, respectively. (Standard errors in parentheses). These are assessments of forecasts of the Democrat's probability of victory. In the encompassing regression with the Democratic sample, the constant $\hat{\alpha} = -0.009$ (se=0.093). In the encompassing regression with the Republican sample, the constant $\hat{\alpha} = 0.235$(se=0.103). A few of the 311 observations are dropped in each sample, because there is no intention for the given party.

VIII. Efficient Poll-Based Forecasts from Secondary Dataset

Our secondary dataset consists of any poll from around the world that we could find that asks both an intent and expectation question. The polls are a hodgepodge of

USA and non-USA elections, executive and legislative offices, and the full range of national elections to smaller districts.

Table 9 summarizes the forecast performance of our two questions in forecasting the winning candidate, mimicking the input in Table 2. Again, we use a very coarse performance metric, simply scoring the proportion of races in which the candidate who won a majority in the relevant poll ultimately won the election. The expectation question is correct more often than the intent question in both 0 to 90 and 90 to 180 days before the election; this difference is statistically significant. Indeed this difference is largest in the 90 to 180 day segment. For reasons we are not quite sure of, the expectation question is essentially random after 180 days, in the data in our dataset, but many of those polls are actually a full year and beyond before the election.

Table 9: Comparing the Accuracy from Secondary Dataset

	Days Before the Election ≤ 90				90 < Days Before the Election ≤ 180				Days Before the Election > 180			
	Proportion of observations where the winning candidate was correctly predicted by a majority of respondents to:											
	Expect	Intent	# Obs	# Elections	Expect	Intent	# Obs	# Elections	Expect	Intent	# Obs	# Elections
President	89.4%	80.7%	161	19	69.2%	61.5%	39	12	59.6%	57.7%	52	11
1936 Electoral College	72.3%	80.9%	47	47	-	-	0	0	-	-	0	0
Governor	78.9%	78.9%	19	9	83.3%	50.0%	6	6	100.0%	100.0%	2	1
Senator	81.8%	90.9%	11	7	-	-	0	0	-	-	0	0
Mayor	100.0%	100.0%	4	2	100.0%	66.7%	3	1	-	-	0	0
Other	80.0%	80.0%	10	9	100.0%	66.7%	3	2	50.0%	50.0%	2	2
USA Total	84.9%	81.3%	252	93	74.5%	60.8%	51	21	60.7%	58.9%	56	14
AUS (Parliament)	88.9%	41.7%	36	3	66.7%	33.3%	21	3	24.4%	66.3%	86	2
GBR (Parliament)	85.0%	90.0%	20	9	100.0%	92.3%	13	7	69.4%	62.9%	62	9
FRA (President)	60.9%	56.5%	23	4	40.0%	20.0%	5	3	-	-	0	0
Other	71.4%	71.4%	7	6	0.0%	0.0%	1	1	0.0%	0.0%	1	1
non-USA Total	79.1%	59.3%	86	22	72.5%	50.0%	40	14	43.0%	64.4%	149	12
Total	83.4%	75.7%	338	115	73.6%	56.0%	91	35	47.8%	62.9%	205	26
Diff (standard error)	7.7%* (2.2)				17.7%* (4.8)				-15.1%* (4.4)			

IX. A Structural Interpretation

At this point, it's worth reflecting on just why it is that the expectation-based forecasts are more accurate. The basic intuition is simply that each of us possesses substantial information that is relevant to forecasting the election outcome, but the voter

intention question only elicits part of this information set. By asking about each respondent's expectation, a poll can effectively aggregate this broader information set.

We can formalize this idea with a very simple model that also does surprisingly well at matching the key facts about voter expectations. We conceptualize a survey respondent's expectation as deriving from her knowledge of her own voting intention, as well as those of $m-1$ of her friends, family, and coworkers, effectively creating a survey of m likely voters. We denote the proportion of person i's sample who plan to vote for the Democrat as $\widehat{v_r^i}$. If each person's individual "informal poll" is drawn from an unbiased sample, then $\widehat{v_r^i}$ is a random variable with mean v_r and variance of $v_r(1-v_r)/m$. Using the normal approximation to the binomial distribution, this suggests that the probability an individual respondent expects the Democrat to win is:

$$Prob\left(\widehat{v_r^i} > 0.5\right) = \Phi\left(\frac{v_r - 0.5}{\sqrt{\frac{v_r(1-v_r)}{m}}}\right) \approx \Phi\left(2\sqrt{m}(v_r - 0.5)\right) \quad [18]$$

where $\Phi(.)$ is the standard normal cdf. The approximation equality follows because in competitive elections $1/\sqrt{v_r(1-v_r)} \approx 2$.

Thus, equation [15] suggests that in a simple probit regression of whether an individual forecasts the Democrat to win, the coefficient on the winning (or negative losing) margin of the Democrat candidate reveals $\sqrt{m}$.[40] The intuition here can be explained by thinking about two extreme cases. If each survey respondent based their expectations on an informal poll of thousands of friends, then even in very close races, most of them will agree on which candidate they expect to win. But if each respondent polls only a couple of friends, then sampling variation will have a larger influence, and there will be much greater disagreement over the likely winner. Thus, the basic logic

[40] We can avoid the approximation noted above, and run a probit regression where the dependent variable is instead $\frac{p_s-0.5}{\sqrt{p_s(1-p_s)}}$; this yields similar results.

identifying our estimate of m can be seen in Figure 2, which shows the likelihood that the expectations of survey respondents are consistent with each other, as a function of how close the race is.

In fact, we have already run the regression described in equation [18]—it is identical to that described in equation [4], which we used to estimate $\widehat{\sigma_e}$. Comparing these equations we see that $2\sqrt{m} = \frac{1}{\sigma_\epsilon}$, or $m = \frac{1}{4\sigma_\epsilon^2}$. Thus given our estimate of $\widehat{\sigma_\epsilon} = 0.150$, we infer that $m = 11$ (the exact coefficient is 11.11, and the standard error clustering by state-year and applying the delta method is 1.12). That is, our simple model suggests that each poll respondent bases her expectation on a poll of herself, plus 10 friends.

We should add a qualifier to this interpretation which highlights that this simple model also serves as a metaphor for a slightly more realistic scenario. In particular, it is likely that the social network of any survey respondent is not a representative random sample. It may be that their social networks contain a known partisan bias or that specific clusters of friends have correlated voting intentions. Our point is simply that the voter expectation question draws upon a broader information set, whose forecasting value is similar to that of an unbiased or random sample with an effective sample size of eleven.

Table 10 shows the relationship between intention, the winning candidate, and expectation in the primary dataset. In approximately half of our individual-level observations (48.8%), the respondent has the same expectation and intention and that candidate wins the elections. In less than 10% of individual-level observations (9.2%) the respondent does not expect the candidate for whom she intends to vote to win, but the candidate does win. In the remaining 42% of observations, where the voter intention is not for the winning candidate, the respondents expect the winner 19.7% and expect the candidate they intend 22.3%. Another way to look at the data is that in 71.1% of

observations the respondent expecting their intended candidate to win: 48.8% correctly and 22.3% incorrectly. And 28.9% of observations the respondent expects their intended candidate to lose: 19.7% correctly and 9.2% incorrectly.

Table 10: Relationship between Voter Intention, the Winner, and Expectation

Intention Democratic	Winner Democratic	Expectation Democratic	Percent of Respondents
1	1	1	19.0%
1	0	1	15.8%
1	1	0	3.4%
1	0	0	12.3%
0	1	1	7.5%
0	0	1	5.8%
0	1	0	6.5%
0	0	0	29.8%

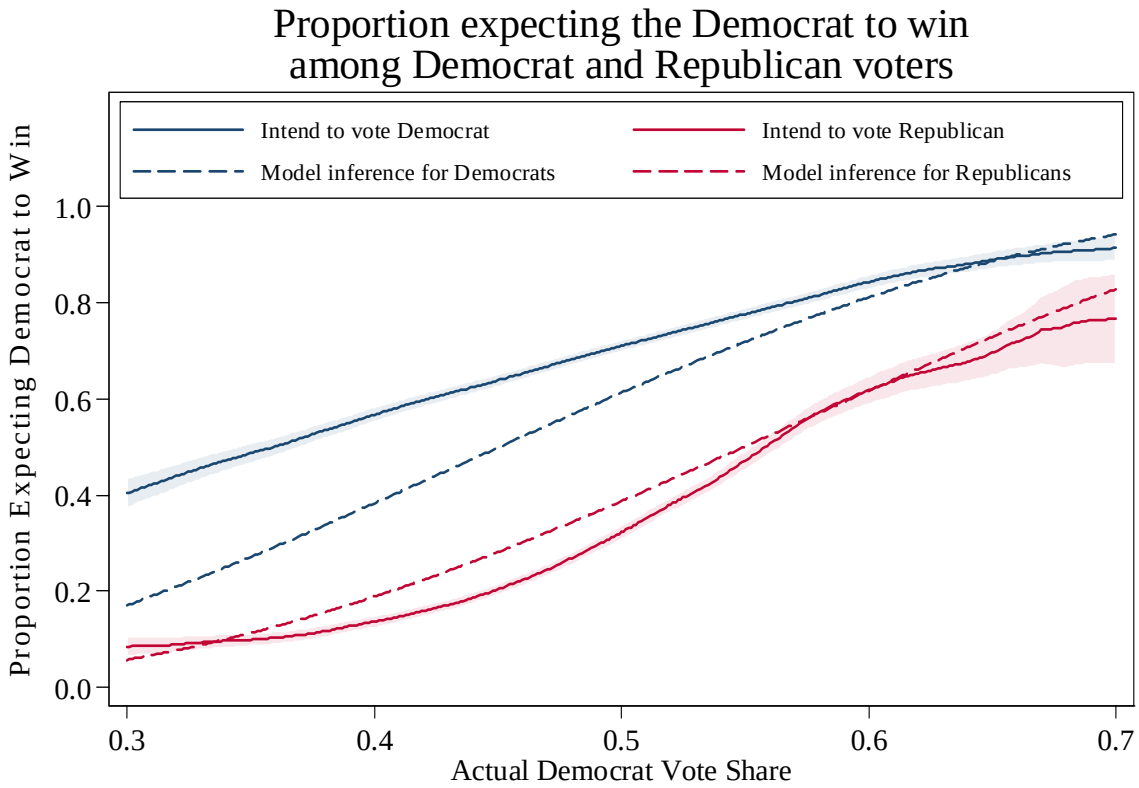
Notes: There are n=13,208 individual-level observations, including 2008 data.

Figure 8 charts the proportion of respondents that expect the Democratic candidate to win for any given actual vote share. We do so for both the full data set of 13,208 individual responses and the structural interpretation from above. From equation [18], if a random respondent has a probability of $\Phi\left(2\sqrt{11}(v_r - 0.5)\right)$ of expecting the Democratic candidate to win, a Democratic supporter has a $\Phi\left(2\sqrt{10}(v_r - 0.45)\right)$ probability of expecting the Democratic candidate to win. One of the eleven votes is guaranteed to the Democrats, so that only 5 of the remaining 10 votes need to go Democratic for the respondent to expect the Democratic candidate to win. Likewise, a Republican supporter expects the Democratic candidate to win with a probability of $\Phi\left(2\sqrt{10}(v_r - 0.55)\right)$.

Table 10 and Figure 8 confirm that there is a mix of information reaching the respondents about the election. As in Figure 2, Figure 8 shows a positive correlation

between vote share and expecting to win, for both types of intended voters. If the respondents only considered their own intention, they would expect their own intended candidate 100%, but only do so 71.1% (Table 10) and there would be a straight line of circles at 0 for the Republicans and 100 for the Democrats (Figure 8). If the respondents only considered an unbiased signal, respondents of different intentions would be equally likely to expect the other candidate, at a given vote share. Figure 8 illustrates that respondents are much more likely to expect a candidate to win, at any given vote share, if they intend to vote for that candidate. We know that there is a noisy signal; 9.2% of respondents switch their expectation away from their intended candidate to the losing candidate (Table 10). Further, the noisy signal cannot be too biased towards the voter's intention, because if all signals were completely biased towards their intention then there would be no explanation for a respondent to expect a losing candidate to win, despite a preference for that candidate. Finally, the figure shows that our simple structural interpretation is a reasonable fit for the actual data.

Figure 8: Relationship between Intention, Expectation, and Actual Vote Share



Local linear regression estimates, using Epanechnikov kernal and rule-of-thumb bandwidth.
Shaded area shows 95% confidence interval.

X. Discussion

A Fox News poll, in the field on September 8 and 9 of 2008, prompted a headline trumpeting McCain's lead in the intent question, but buried Obama's continued lead in the expectation question.[41] This is an editorial choice that almost any polling or news organization would have made. Yet, while McCain and Obama occasionally traded the lead in the intent question, Obama, the eventual winner, led in every expectation question we could find. We hope that this paper will give editors, as well as pollsters, campaigns, and researchers, pause in future elections, as we have proved the expectation question significantly useful in creating accurate forecasts of all major outcome variables.

[41] http://www.foxnews.com/story/0,2933,420361,00.html

The structural interpretation of the response to the expectation question helps illustrate why expectations are such a powerful polling tool. The answer we receive from the expectation question includes all of the information in the intention question as well as the intention of approximately ten friends, family, and coworkers who are likely voters. This multiplication of the sample size is crucial in making the expectation question remarkably valuable with small sample sizes. That phenomenon, combined with the identification of likely voters, is what enables expectation polls to provide valuable information, even with a non-random sample of respondents.

The structural interpretation offers clues into the ratio of the sources of data that people are using in creating expectations of elections and this can be applied to a wide set of disciplines. One example is questions about economic voting. There is very important question about whether people vote (or frame their political preferences in general) on their own pocketbook (egotropic) or the national economy (sociotropic). We have a tool to look at which source of information is framing their expectations, where their expectations are the main input in their preferences or revealed utility. The same example can be done for the popularity of a new style of jeans in marketing or for consumer sentiment on durable goods in coming months in economics. In any situation where individual-level stakeholders are exposed to private information and public signals the results and methods of this paper illustrate meaningful information that can be extracted by individual-level questions of expectations.

XI. References

Alford, Richard F. 1977. "Estimation of the Probability of Bryan Victory in 1896" Presented at the 1977 *Cliometrics Conference.*

Barreto, Humberto and Frank Howland. 2006. *Introductory Econometrics: Using Monte Carlo Simulation with Microsoft Excel*. New York, NY: Cambridge University Press.

Berg, Joyce E., and Thomas A. Rietz. 2006."The Iowa Electronic Market: Stylized Facts and Open Issues.". *In:* Hahn *Robert W.,* Tetlock *Paul, editors. Information Markets: A New Way of Making Decisions in the Public and Private Sector, eds. Robert W. Hahn, and Paul Tetlock.* Washington, DC*:* AEI-Brookings Joint Center*.*

Berg, Joyce E., Forrest D. Nelson and Thomas A. Rietz. 2008. "Prediction Market Accuracy in the Long Run," *International Journal of Forecasting* 24(2):283-298.

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

Erikson, Robert S., and Christopher Wlezien. 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.

Granberg, Donald and Edward Brent. 1983. "When Prophecy Bends: The Preference-Expectation Link in U.S. Presidential Elections." *Journal of Personality and Social Psychology* 45(3):477-91.

Lock, Kari and Andrew Gelman. 2010. Bayesian Combination of State Pols and Election Forecasts. *Political Analysis*, 18(3):337-348.

Imai, Masami and Cameron Shelton. 2010. "Elections and Political Risk: New Evidence from Political Prediction Markets in Taiwan." Wesleyan Economics Working Papers.

Irwin, Galen and Joop Van Holsteyn. 2002. "According to the Polls, the Influence of Opinion Polls on Expectations." *Public Opinion Quarterly* 66:92-104.

Moulton, Brent. 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Unites." *The Review of Economics and Statistics* 72(2):334-338.

Mutz, Diana. 1995. "Effects of Horse-Race Coverage on Campaign Coffers: Strategic Contributing in Presidential Primaries." *The Journal of Politics* 57(4):1015-1042.

Robinson, Claude E.. 1937. "Recent Developments in the Straw-Poll Field" *Public Opinion Quarterly* 1(3):45-56.

Rothschild, David. 2009. "Forecasting Elections, Comparing Prediction Markets, Polls, and Their Biases." *Public Opinion Quarterly* 73(5):895-16.

Wolfers, Justin and Eric Zitzewitz. 2004. "Prediction Markets." *Journal of Economic Perspectives*, 18(20 107-126.

Expectations: Point-Estimates, Probability Distributions, Confidence, and Forecasts

Abstract

*In this article I t*est a new interactive web-based interface that captures both "best estimate" point-estimates and probability distributions from non-experts. As in the previous literature, respondents are overconfident. My innovation is to show that in contrast to standard methods of directly asking respondents to state their confidence, using my method, which induces the respondents to reveal confidence, there is a sizable and statically significant positive relationship between confidence and the accuracy of individual-level expectations. This positive correlation between confidence and accuracy can be utilized to create confidence-weighted aggregated forecasts that are more accurate than the standard "consensus forecasts." The payment of financial incentives does not affect these findings.

I. Introduction

The main method tested in this article captures "best-estimate" point-estimates and then probability distributions from non-experts on upcoming events. Utilizing a graphical, interactive interface, only possible with web-based polling, my method provides several innovations to ease the creation of probability distributions by non-expert respondents. The point-estimates and probability distributions are used to explore both the nature of individual-level expectations and forecasts, especially in regard to the individual-level confidence, as revealed through the size and shape of the probability distribution.

The responses demonstrate a slight overconfidence. In testing confidence, I compare my method of gathering probability distributions to the two most attractive

alternatives: a simple four point scale of stated confidence and the best calibrated/transparent method of asking for a confidence range around a point-estimate. Where comparable, the probability distributions and confidence range provide a similar slight overconfidence; this overconfidence is consistent with the literature (Soll and Klayman, 2004; Teigen and Jorgensen, 2005; and Speirs-Bridge et al., 2010).[42]

There is a sizable and statistically significant correlation between confidence from the probability distributions and the accuracy of the corresponding expectations that has not been previously documented (where confidence is estimated as the inverse of the variance). There is variation in the levels of confidence both within questions (i.e., between respondents) and within respondents (i.e., between questions), and this correlation holds for both situation. This correlation is weaker and/or non-significant in the two simplified methods. Engleberg et al (2009) determines that the distribution of probabilities widens as the event horizon lengthens. Yet, that article does not address the correlation with the size of a distribution and the accuracy for a given event. Other articles conclude that tighter confidences ranges correlate with increased overconfidence (i.e., at best, confidence is uncorrelated with accuracy for non-experts).[43]

Confidence can be used as a tool for weighing in the creation of aggregated forecasts of the level of the outcome; these confidence-weighted forecasts are more accurate than forecasts from the point-estimates alone. There is consensus that aggregating expectations creates more accurate forecasts, on average, than picking individual-level expectations; Pennock and Reeves (2007) demonstrates that with a self-selected group of respondents making NFL predictions only 6 of the 2231 players are

---

[42] Confidence is comparable in the percentage of answers that come to fruition within a confidence range (a set of perfectly calibrated 80% confidence ranges would have 80% of the answers occur within the range). Stated confidence cannot provide a measure and the confidence range provides only the one range it was asked, where the probability distribution reveals an infinite set of confidence ranges.

[43] Kuklinski (2000) concludes that for objective political questions confidence can actually be negatively correlated with accuracy.

more accurate than the collective average. Further, there are numerous studies showing that the simplest aggregation techniques are the most efficient (Bates and Granger, 1969; Stock and Watson, 2004; Smith and Wallis, 2009). The simplest weight considered in these articles, other than an average, is the inverse of the mean square error; this occasionally provides a slightly smaller error than averages. The questions for this article are answered concurrently by non-experts, providing no opportunity for historically derived weights; yet, if inverse of the variance of their probability distributions is well calibrated, it could serve as a proxy for mean square error.[44]

This article addresses various literatures where Likert-type rating scales are used to supplement estimates. Phycology, political science, and marketing are just a few literatures where there are standard survey practices that ask respondents to state their: confidence, likelihood, or agreement with their previously stated estimations. Generally these are ordered, one-dimensional scales that are designed to provide an extra dimension to the estimation. In Likert-type rating scales, responding to varying incentives, some respondents gravitate towards or away from extreme choices, regardless of their true state, using a revealed method for confidence, respondents are unaware of the extremities of reasonable answers or provided any reference if they wish to manipulate the strength or weakness of their response. With Likert-type rating scales respondents see these questions as separate follow-up questions, rather than the main answer, providing different levels of attention and incentives, this article tests methods of having the respondents reveal their confidence inside the main question. Finally, the graphical nature of the method allows the researcher to subtly provide information that allows respondents to answer more complicated questions revealing their confidence, where the questions would be very difficult to ask in a telephone or in-person setting. Thus, the revealed confidence from my method is more representative of the

[44] Thus, I do not have the data to attempt updating on risk profile as in Chen et al (2003) or more complex Bayesian methods like those cataloged in Clemen (1989) or suggested in Clemen and Winkler (1993).

respondents' information, rather than their natural state of confidence or a range manipulated for any other incentive than the "best estimate" of the respondent.[45]

Prediction markets weigh their individual-level data by how much money people are willing to gamble, a proxy for confidence; this article illustrates the benefits of a new method of directly weighing individual-level polling data by confidence. There is growing consensus that prediction markets aggregate individual-level expectations into more accurate forecasts than standard polling methods. Rothschild (2009) confirms the accuracy of prediction market-based forecasts and outlines some of the differences between them and standard polls as methods of capturing and aggregating individual-level responses: the nature of their sample of respondents, the question asked to the respondents, the aggregation method of the individual-level data, and the incentives for the respondents. Rothschild and Wolfers (2011) addresses the question being asked of the respondents; that article concludes that eliciting expectations, as prediction markets do, captures more information from the respondent than the questions asked in standard polls. This article addresses another of the differences between prediction markets and polls, the weighting of the individual-level data.

In this article I test a graphical, interactive web-based interface to advance the ability of researchers to collect and utilize individual-level expectations. I capture both point-estimates and probability distributions from non-expert respondents. The results provide insight into expectations, revealing the nature of their: accuracy, confidence, and calibration of probabilities. Researchers can use these expectations as individual-level data for aggregated forecasts and, in the future, to understand heterogeneity in revealed behavior under uncertainty. Better understanding of non-expert expectations

[45] Barber and Odean (2001) shows that men are more overconfident than women, as demonstrated with more aggressive behavior in the stock market, regardless of knowledge. My method reveals the confidence range, without the other confounding factors of utility in a decision such as investing. One reason for the results in Kuklinski (2000) is that respondents perceive an incentive to be strategic about stated confidence levels in political questions, but this is not an issue when people are revealing, rather than stating their confidence.

will allow researchers to learn more about the absorption and transformation of information. Better forecasts help researchers connect shocks with changes in the underlying values of the world and investors make more efficient use of their time and money. As telephone-based polling is supplanted by web-based polling, I hope that this article will inspire further exploration into graphical, interactive web-based interfaces that can ask questions and consequently gather information not possible in standard telephone-based or even in-person-based polling.

II. Method

I build on the most recent methods of surveying expectations to create a graphical, interactive web-based interface that gathers expectations: point-estimates and probability distributions. One influence on my method is Delavande and Rohwedder (2008), which asks respondents for a point-estimate (on how much social security they will receive) and then use a visual screen that asks the respondents to distribute 20 balls (each representing a 5% probability of the outcome) in a series of bins that signify possible value-ranges of the outcome. I enhance their method with lessons from the literature involving the generation of confidence ranges around point-estimates. A few key innovations of my method: it distributes probabilities as small as 1% into upwards of nine bins of value-ranges, forces the respondent to consider the question from multiple angles, and uses new graphical tools to efficiently clarify the procedure for the respondent. Further, after I collect these expectations (i.e., the point-estimate and probability distribution) I probe other characteristics of the respondents that may be correlated with biases or varying information levels.

The first piece of data I recover from the respondent for any specific question is a "best estimate" point-estimate. While maintaining overall consistency regardless of the specifics of the questions, the interface is adjusted enough to ensure that the respondents provide valid responses. The appropriate interface allows the respondent to understand

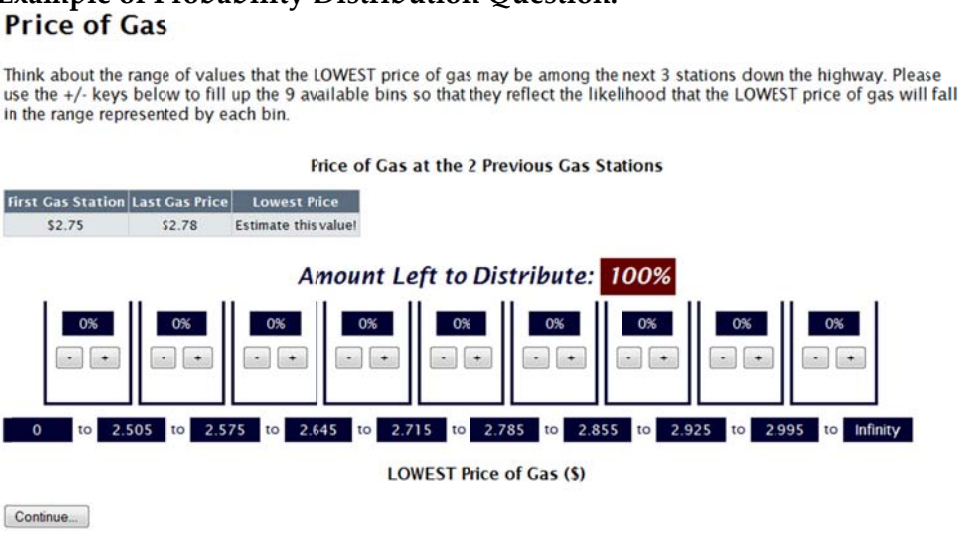
exactly what the question asks, without taking up too much time and effort. Further, it takes measures to avoid any anchoring or suggestive examples for the respondent. As an example, the graphical design for vote share questions uses a slider that shows the two-party vote share for both candidates, which makes it easy to understand the basis of two-party vote share.

Example of Point-Estimate Question:

The second piece of data elicited is a probability distribution. A non-interactive poll would ask a series of questions to the respondent to create a distribution of probabilities or it could provide a series of pre-set value ranges for the respondent to distribute her probabilities. My method creates a series of value ranges centered on their point-estimate. The respondent distributes "100 likelihoods" into the 9 different bins (each representing a value range). Further, if the ranges prove too wide and the respondent places more than 50% of the probability in one bin, the program creates a new set of bins within the aforementioned range. The respondent must answer fully and cannot be internally inconsistent (i.e., the overall probability must equal 100% and there cannot be contradicting probabilities), thus all responses can be used. By adjusting the bins' ranges so that they are centered on the point-estimate, the method eliminates irrelevant thresholds and anchoring derived from pre-determined thresholds.

Respondents can answer more questions faster, because instead of a series of questions, there is just one question to get a distribution. Consistent with the literature, I assume that probabilities are distributed uniformly within a bin.

Example of Probability Distribution Question:



There are two additional sources of data from the respondents. First, before the first task, the respondent is asked a series of questions on observable characteristics. Second, within each question, the two main questions are followed by a question which examines the respondents' personal partiality related to the question. As an example, if the questions were about movies, I would ask the respondent about their intention to see the movie.

The data for this article was gathered in two rounds of studies. In both of the studies there were five categories of questions, with each category having either 9 or 10 unique questions. Each respondent was in just one of the studies and answered just one question from each category. Everything is randomized so that each respondent sees the categories in a random order, is randomly assigned to a unique question in a category, and, if there is variation in the information level for a given question that is also randomly assigned. Both studies are a mixture of two groups of people, the Wharton

Behavioral Lab (mainly students and staff) and respondents from around the USA with Mechanical Turk.

Study I concentrates on comparing my method to the methods of stated confidence and confidence intervals. Each respondent answered one question in the following five categories: calories, concert ticket prices, gas prices, movie receipts, and unemployment rate (for question details, see Appendix A). Half of the respondents were randomly assigned to create the full probably distribution and half were asked state their confidence and create a confidence interval. The stated confidence is recovered with the standard polling question: "How confident are you of your answer?" with a drop down menu of four choices: very confident, somewhat confident, not very confident, and not at all confident. The confidence range is recovered in the most efficient method, asking the respondent the combination of: "I am 90% sure answer is greater than _____" and then "I am 90% sure the answer is less than _____". Soll and Klayman (2004) demonstrates that respondents reduce overconfidence when they are asked to formulate an upper and lower limit in separate steps. The theory is also demonstrated in Vul and Pashler (2008); that article shows that when respondents are forced to think about the same question in two different ways, they add new information that is not apparent when they just consider the question from one angle.[46]

Study II concentrates on comparing my method under different incentives; this study responds to questions from Becker, Degroot, and Marscak (1964), which

[46] Teigen and Jorgensen (2005) concludes that overconfidence is decreased when respondents assign probabilities to confidence ranges, rather than confidence ranges for a set probability and similarly Speirs-Bridge et al. (2010) argues that it is best to ask the respondent to re-calibrate his own range (i.e., show them their 80% range they just created and ask them how wide that really is). The full probability distribution method requests the respondent to input their own probability to set ranges, rather than input a range to a set probability, but that is not possible with the confidence ranges. I do not want to make ungrounded assumptions on how to translate a 60% range into an 80% range. And, without trusting their assumptions and generating uniform ranges, there is no method of comparing the respondents' on their calibration, as they would all have different size ranges.

determines that any elicitation of expectations should be incentive compatible. Each respondent answered one question in each of the following five categories: calories, gas prices, and unemployment rates (as in Study I), and then voter turnout and Senate election results (for question details, see Appendix A). The voter turnout and Senate election results are done with the respondents' home states, so due to lopsided draws, are not included in this article, but the basis for further research. Movies and concerts are dropped as categories, because Study II was done in the early fall, where the time frame was too short for movies and concerts are not frequent enough. While the first three categories are the same, all of the questions are new. Half of the respondents are paid with standard flat fee and half of the respondents are paid a flat fee and then incentivized with an incentive compatible scoring rule. The five respondents with the lowest weighted square error over all of their responses were given the bonuses.[47] Since Study II is unbalanced in its categories, I only use it in the full data when I am making comparisons concerning the effect of the incentives.

III. Estimation/Results

Between-respondent disagreement is much larger than within-respondent uncertainty; this demonstrates both the overconfidence of the responses and issues involving the accuracy of the stated point-estimates. This comparison is illustrated in coefficients of variation in Table 2, where between-respondent disagreement is the coefficient of variation of the stated point-estimates for a unique question and within-respondent uncertainty is the average coefficient of variation of the individual-level probability distributions for that unique question. Gurkaynak and Wolfers (2006) studied an economic derivative market and show that between-respondent

[47] I note in parenthesis in the directions "i.e., the most accurate distributions", because very few respondents are going to know what a mean square error is, or what type of response minimizes it. This inability to comprehend the scoring rule is a problem for incentives noted in Artinger (2010).

disagreement of point-estimates (submitted by experts) is less dispersed than within-forecast uncertainty, illustrated by an efficient market. One reason that the between-respondent disagreement is relatively larger, is that in this article I am not studying the uncertainty in an efficient market, but within each individual, where the standard deviation of individual-level probability distributions reflects individual uncertainty of the outcome. In two paragraphs I will demonstrate that these probability distributions are too narrow (i.e., overconfident). The second reason is because the respondents are providing point-estimates that extend all over their distributions, not just the mean or median. The standard deviation of the point-estimates for one question should be similar to the standard deviation of the most likely outcome perceived by the respondents of that question. Yet, for some categories of questions the average absolute log difference between the mean and median of a respondent's distribution and their point-estimate approaches 10%.[48] Table 4 provides further insight into the accuracy of the stated point-estimates.

Table 1: Coefficients of Variation of Individual-Level Probability Distributions and Coefficients of Variation of Point-Estimates

	Study I		Study II Incentivized / non-Incentivized	
Category	**Uncertainty**	**Disagreement**	**Uncertainty**	**Disagreement**
Calories	0.221	0.373	0.175/0.183	0.392
Concert Tickets	0.222	0.384	-	-
Gas Prices	0.026	0.015	0.027/0.026	0.027
Movie Receipts	0.314	0.549	-	-
Unemployment	0.013	0.039	0.018/0.018	0.095

Note: Study I is 120 respondents. Study II is 103 respondents non-incentivized and 99 respondents incentivized. Coefficient of variation for uncertainty is (standard deviation)/(mean of distribution) and for disagreement is (standard deviation)/(mean of point-estimates).

The data from Study II shows that the alignment of the incentives has a negligible influence the individual-level distributions. The incentive compatible pay

[48] While very few point-estimates that occur on the tails of the probability distributions, for calories, concert tickets, and movie receipts the mean or median is larger than the point-estimate by a statistically significant amount.

rewarded respondents extra for properly calibrated distributions. Table 1 demonstrates that the coefficients of variation are very similar, regardless of incentives. Regardless of the outcome, incentives are not ideal for this project, so it is comforting they have a negligible impact. First, if the researcher outlines the type of response that maximizes the payout rule, she is manipulating the response, but, in this project, I want the response to be whatever the respondent thinks is the "best estimate" not what I define as the "best estimate". Further, in future articles in this project where I connect the expectations to decisions, I want the true expectations, not expectations created to fulfill my scoring rule. Second, non-payment or flat fees are standard in polling and I want the results of this project to be relevant.

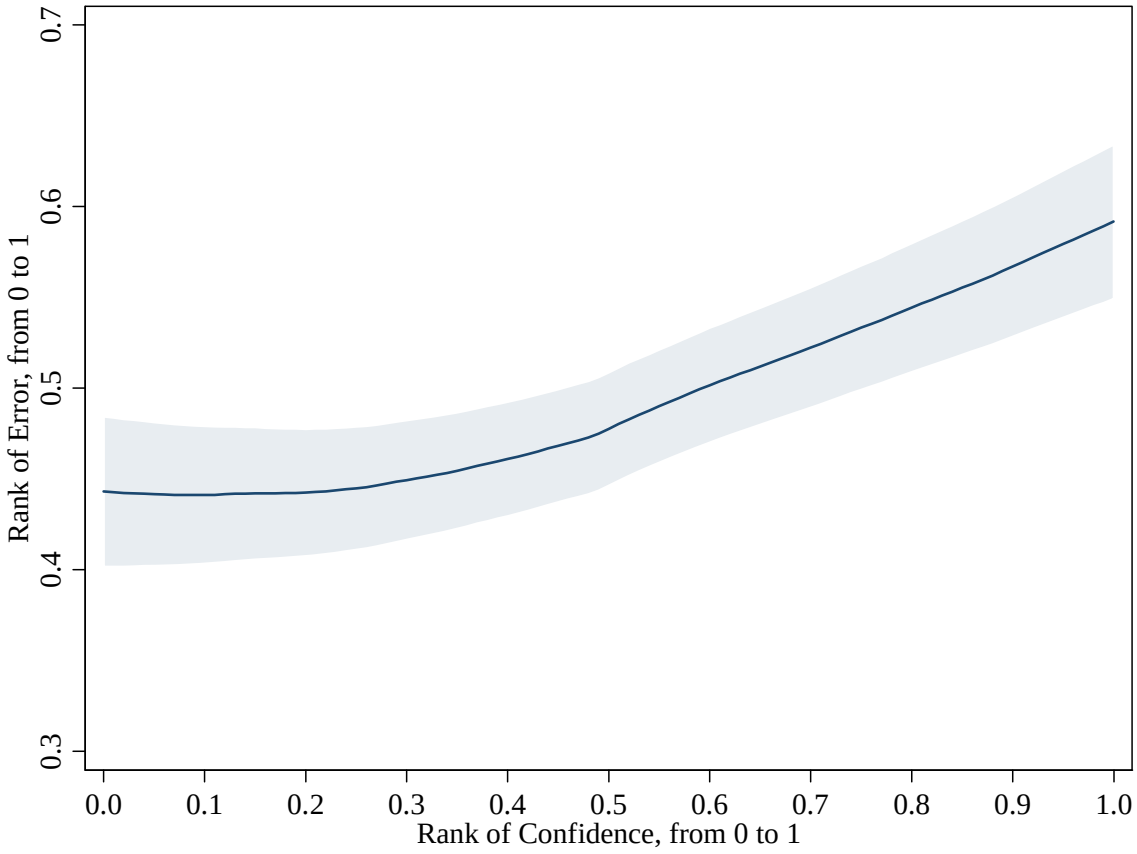
The calibration of confidence for the individual-level probability distribution is very similar to the confidence ranges (where it is comparable); both are slightly overconfident. Utilizing the data from Study I, the answer lies within the 80% confidence range 60% of the time and within the middle 80% range of the probability distribution 58% of the time, a statistically insignificant difference. These are both in line with most comparable studies, where non-experts were asked simple knowledge questions (e.g., the ranking of college or the winning % of NBA teams).[49] The probability distribution forces the respondent to consider the answer for more time, but the confidence range implores them to consider the question from more angles. This systematic overconfidence is not a problem for this article, as I care more about the relationship between confidence and accuracy.[50]

Confidence derived from the probability distributions correlates positively with accuracy; this correlation is more sizable than the correlation from the confidence ranges, and the correlation from the stated confidence is positive, but statistically

[49] Soll and Klayman (2004), Teigen and Jorgensen (2005), and Speirs-Bridge et al. (2010).
[50] Again, in future work the expectations will be used to inform decision making, so I prefer authentic representation of the expectations relative to forcing more accurately calibrated confidence ranges.

insignificant. In order to make comparisons across the different types of data and the different categories, I simplify the data. Within each question I rank the responses from the least to most confident according to their standard deviation, width of confidence interval, or stated confidence from 0 to 1 on its position relative to all of the responses to its unique question. The smallest standard deviation, narrowest confidence range, or most confident answer is ranked 0, where the highest, widest, and least confident is ranked 1. I do the same for the error of the point-estimate. Figure 1 illustrates the relationship between confidence and accuracy in the probability distribution data. Relatively low confidence correlates with a lower than average, average rank of error, and relatively high confidence correlates with a higher than average, average rank of error. Table 2 shows that the correlations are positive and significant within unique questions for both the confidence range and probability distribution, but nearly twice as sizable for the probability distribution. The correlation is positive, but not significant, for stated confidence.

Figure 1: Correlations Between Confidence Derived From Probability Distributions and Accuracy of Point-Estimate



Note: Local linear regression estimates, using Epanechnikov kernel and rule-of-thumb bandwidth. Shaded area shows 95% confidence interval.

Table 2: Correlations Between Confidence and Accuracy of Point-Estimate

	Stated Confidence	Confidence Range	Probability Distribution	R^2
$\boldsymbol{Rank(Error) = \alpha + \beta * Rank(\sigma)}$	0.035 (0.038)	-	-	0.000
	-	0.151*** (0.040)	-	0.023
OLS (Within Question)	0.006 (0.038)	0.150*** (0.041)	-	0.023
	-	-	0.231*** (0.040)	0.053

	0.103** (0.050)	-	-	0.001
$\boldsymbol{Rank(Error) = \alpha + \beta * Rank(\sigma)}$	-	0.233*** (0.051)	-	0.023
Fixed-Effect (Within Respondent)	0.070 (0.050)	0.222*** (0.052)	-	0.022
	-	-	0.260*** (0.052)	0.053

Note: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10% level, respectively. (Standard errors in parentheses). The errors and standard deviations are normalized by their rank within the unique question. The stated confidence and confidence range questions were answered by 129 respondents and the probability distribution by 120. There are a total of 48 unique questions in 5 categories; each respondent answered 5 questions, one in each category.

Confidence is not only correlated with accuracy within questions, but within respondents. Respondents adjust their confidence, relative to the other respondents, depending on their confidence for a given question. It is relatively easy for a respondent to shift her confidence between questions when she is using a confidence scale, as the values remain constant between questions. Yet, for a confidence range or probability distribution, the relative confidence is uncertain, as average range sizes and standard deviations vary widely between questions categories. This inability to consciously gauge relatively confidence makes it difficult for respondents to manipulate their confidence response for any other incentives than their truthful response.

The wisdom of the crowd is working; the mean or median of the question's responses is more accurate than a random respondent for the majority of answers and has a significantly smaller error. This is shown in Table 3. For example, in Study I, just 24.3% of respondents' point-estimates are more accurate than the median point-estimate of the respondents. Further, the error for the median is less than that of the mean, on average and for the vast majority of questions. It is likely that the outliers are bigger issue in non-expert point-estimates than in expert point-estimates; either way, the

median is a standard point-estimate to use for aggregating point-estimates into forecast.[51]

Table 3: Individual-Level Point-Estimates

	Study I	Study II
Categories	5	3
Questions per Category	9.6	10
Observations per Question	25.8	20.1
% of Individual-Level Point-Estimate Absolute Errors < Mean Point-Estimate of Question Absolute Errors	36.7 %	38.8 %
% of Individual-Level Point-Estimate Absolute Errors < Median Point-Estimate of Question Absolute Errors	24.3 %	27.9 %

Note: Point-estimates are all recorded prior to the probability distributions. Study I is randomized between probability distribution method and confidence questions, with 249 respondents. Study II is randomized between flat pay and incentive compatible pay for probability distribution method, with 202 respondents.

Point-estimates derived from the probability distributions are more accurate than the stated point-estimate. I test three simple point-estimates from the probability distributions: mean, median, and mode. On the top of Table 4 I run a Fair-Shiller (1989 and 1990) regression, which includes the point-estimate and the probability distribution's point-estimates, with fixed-effects by category of question. When compared directly with the stated point-estimates, the mean and the median of the probability distribution are both significant at the 10% level where the stated point-estimate is not significant, and the coefficients for the mean and median are substantially larger than the stated point-estimate's. I cannot rule out that the mode provides no information that is not in the stated point-estimate. By switching to OLS with no constant and constraining the coefficients to sum to 1, I can determine the optimal weights of the different variables if I was forced to put them together for a best estimate. Again, compared directly, the mean and the median are both significant and the stated point-estimate is not. There are two plausible explanations for this finding. First, I

[51] Galton (1907) recommends the median for non-experts guessing the weight of a cow (a point-estimate) and this is the inspiration shown repeatedly in Surowiecki's *Wisdom of the Crowd*. Engleberg, et al (2009) also uses the median for GDP and inflation with experts.

cannot rotate the order of the stated point-estimate and the probability distribution, thus the respondent may be making a more accurate estimate in the probability distribution versus the stated point-estimate, because it is her second chance to consider the question. Second, the respondents may be ignoring long asymmetric tails of the probability distribution when they state their point-estimates, to the detriment of their accuracy of their stated point-estimate.

Table 4: Comparing the Point Accuracy of the Individual-Level Expectations

$DistVar$ in regression:	Mean	Median	Mode	Point-Estimate
$ans = \alpha + \beta_1 DistVar + \beta_2 PointEst$	0.311*	-	-	0.187
	(0.181)			(0.179)
	-	0.287*	-	0.210
		(0.170)		(0.168)
	-	-	0.018	0.472***
			(0.142)	(0.147)
	-0.310	0.824	-0.314	0.307
	(0.824)	(0.840)	(0.221)	(0.198)
$ans = \beta DistVar + (1-\beta) PointEst$	67.8%***	-	-	32.2%
	(22.3)			(22.3)
	-	61.9%***	-	38.1%*
		(0.211)		(0.211)
	-	-	-0.228	1.228***
			(0.179)	(0.179)
	-1.423	2.872***	-1.211***	0.763***
	(0.962)	(1.020)	(0.267)	(0.241)

Note: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10% level, respectively. (Standard errors in parentheses). There are fixed-effects by category in top regression. If I run separate OLS regressions with a constant on each category: the mean has a much larger coefficient (and more significance) in 3 categories, the point-estimate in 1 category, and similar in 1 category. If I run a regression for each expectation type by itself, the mean has the highest R^2. There are total of 590 observations in the five categories.

Weighing the point-estimates by confidence produces more accurate forecasts of the outcome level than other standard methods. Table 3 illustrates how without a probability distribution the most accurate consensus forecast of the outcome level is the median of the point-estimates for any question. Table 4 shows that on an individual-

level, the mean of the probability distribution is the most informative of its point-estimates (and more informative than the median of the stated point-estimates). Following the literature's use of the inverse of the mean square error, I use the simplest and most transparent comparable method of aggregating the means of the probability distributions to create a consensus forecast; I weigh each mean by the inverse of their standard deviation squared (i.e., the variance):

$$w_i = \frac{1/\sigma_i^2}{\sum_{j=1}^{n} 1/\sigma_i^2}$$

Any response, i, has the weight of $1/\sigma_i^2$ divided by the sum of all of the inverse variances of the responses to its unique question. This method is efficient only if the responses are efficient, but I have already shown they are overconfident. Yet, this is the most transparent and universal method available, and it does provide a more accurate answer.

Illustrated in Table 5, the confidence-weighted forecasts have more weight and/or significance more often than the median of the point-estimates. There is certainly fluctuation between the categories, but the confidence-weighted mean demonstrates itself to be meaningful in relation to the standard method.

Table 5: Comparing the Point Accuracy of the Most Promising Forecasts

Category	Weight	Median of Point-Estimate	Confidence-Weighted Mean	Median of Point-Estimate	Confidence-Weighted Mean
	$ans = \alpha + \beta_1 PointEst + \beta_2 ConEst$			$ans = \beta PointEst + (1-\beta) ConEst$	
Calories	$1/\sigma_i^2$	0.059	1.146***	0.052	0.948***
		(0.286)	(0.281)	(0.245)	(0.245)
Concert Tickets	$1/\sigma_i^2$	0.730	0.282	0.390	0.610
		(0.822)	(0.677)	(0.564)	(0.564)
Gas Prices	$1/\sigma_i^2$	-0.315	-0.021	-0.405	1.405
		(0.398)	(0.425)	(1.133)	(1.133)
Movie Receipts	$1/\sigma_i^2$	0.805**	-0.791*	0.458	0.542
		(0.319)	(0.348)	(0.453)	(0.453)
Unemployment	$1/\sigma_i^2$	-1.052	2.097	-0.480	1.480
		(1.786)	(1.808)	(1.553)	(1.553)
	$ans = \alpha + \beta_1 PointEst + \beta_2 ConEst$			$ans = \beta PointEst + (1-\beta) ConEst$	
Calories	$1/\sigma_i^{1.5}$	0.007	1.186***	-0.020	1.020***
		(0.286)	(0.279)	(0.246)	(0.246)
Concert Tickets	$1/\sigma_i^{1.5}$	0.726	0.303	0.272	0.728
		(1.064)	(0.940)	(0.861)	(0.861)
Gas Prices	$1/\sigma_i^{0.3}$	0.342	-0.633	0.345	0.655
		(0.846)	(0.823)	(2.463)	(2.463)
Movie Receipts	$1/\sigma_i^{0.1}$	-0.499	0.792	1.947**	-0.947
		(0.829)	(0.773)	(0.729)	(0.729)
Unemployment	$1/\sigma_i^{2.7}$	-1.017	2.055	-0.738	1.738
		(1.653)	(1.666)	(1.515)	(1.515)

Note: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10% level, respectively. (Standard errors in parentheses). There are 48 question total: 10 for calories, 10 for gas prices, and 10 for unemployment, 9 for concert tickets, and 9 for movie receipts.

In the second part of Table 5 I explore more efficient weighting for the confidence-weighted forecasts. Rather than raising the standard deviation to the exponent two (i.e., squared), I allow the exponent to fluctuate to whatever generates the smallest root mean square error for the predicted forecasts from the confidence-weighted forecast. There would be a different set of exponents if I take the exponents that minimize the mean square error of the raw confidence-weighted forecasts or what minimizes the jointly predicted mean square error of the confidence-weighted forecast

and the median of the point-estimates. Yet, regardless of which set of efficient exponents I use to create the confidence-weighted forecasts, the weighting between the median point-estimate and the confidence-weighted forecast are similar.

Table 6 looks at the R^2 from the regression of the answer on just the median of the point-estimate, on just the confidence-weighted forecast, and on both the median of the point-estimate and the confidence-weighted forecast. For two of the categories, the growth in R^2 from median of point-estimate to having both forecasts is minimal, but in the other three there are substantial gains (i.e., there is explanatory power in the confidence-weighted forecast).

Table 6: Comparing the Point Accuracy of the Most Promising Forecasts, with R^2 from $ans = \alpha + \beta * Forecast$

Category	R^2 with only Median of Point-Estimate	R^2 with only Confidence-Weighted Forecast	R^2 for Joint Forecast
Calories	0.585	0.884	0.884
Concert Tickets	0.880	0.873	0.882
Gas Prices	0.308	0.347	0.362
Movie Receipts	0.131	0.129	0.534
Unemployment	0.985	0.987	0.988

Note: The confidence-weighted forecast is optimized by category as in the lower half of Table 5. The table is nearly identical regardless of which efficient weighting scheme I utilize.

Two categories of questions have follow-up questions that can be used to better calibrate individual-level responses. First, for the calorie question, I ask "How closely do you follow calories?" with options that run from "not very closely" to "regularly, and I know the item in the question." There is no correlation between self-reported information and the rank of absolute error within the question.[52] This conforms to results of Table 5, where similar to stated information, stated confidence levels on a four point scale fails to gain significance in its correlation with accuracy. Second, for the movie

[52] Study I is slightly negative and insignificant and the non-treated Study II is slightly positive and insignificant. Together the coefficient from OLS of intent on rank of error is 0.001 (0.022).

question, I ask "Are you interested in seeing 'MOVIE'?" with options that run from "Definitely not going to watch it" to "Have already been excited about it/Definitely watch it in a theater." Thus, this question is combining two dimensions: information and partiality. There is a positive and statistically significant correlation between information/intent and the point-estimate; if a respondent likes a movie, she projects it to earn more money. A fixed-effect regression of information/intent (which runs from 1 to 5) on the point-estimate yields a coefficient of 19.47 (5.33) for information/intent (i.e., respondents estimate the movie to earn $19.5 million more for each degree of intent to see the movie). With repeated data, I would be able to inflate or deflate responses to debias them for the information and partiality of the respondents.

IV. Discussion

There are several reasons why a graphical, interactive interface can collect information that is difficult to attain in standard telephone or in-person settings. First, information can be revealed, rather than stated, which makes it much harder for the respondents to manipulate the answer to fulfill incentives other than their best estimate. Second, the interface makes the confidence revelation part of the main question, whereas asking a respondent to state confidence after they supply the main point-estimate, may not seem as serious to the respondent or operate under different incentives. Third, Ariely, et al (2003) shows that people can incorporate new information into their understanding of the world; the problem is that they sometimes appear arbitrary, because they are not sure where with what baseline they should start. A graphical interface can provide some subtle baselines for the respondent without providing too much anchoring. For example, in this article, the question regarding calories of fast food includes pictures and descriptions of a few different foods. Similarly, the presentation of the questions themselves are providing subtle information about point-estimates and probability distributions that teach people how to provide information they have, but do not know how to elucidate.

Polls and predication markets are just two methods for gathering individual-level information and aggregating it into forecasts; both methods have benefits and negatives, and my method is one attempt to harness the better aspects of both of them. One of the key problems with polls is the reluctance of researchers to ask the question they are trying to answer, which is usually the question that gathers the most relevant information from the respondent. The graphical and interactive nature of this method allows me to ask questions that do not gather consistent and meaningful responses in a telephone or in-person format. Polls' transparent aggregation does not take advantage of disparities in information of the respondent and prediction markets' more opaque aggregation does not record massive amounts of information and is susceptible to manipulation.[53] With my method I can capture all of the information and aggregate it, transparently, with confidence. Further, I can create not only accurate forecasts of the level of the outcome, but also, I can explore full probability distributions on both the individual and aggregate level.

The full method proves itself meaningful in absolute terms and trumps simpler confidence ranges in information, but it does take up more time, which can be important in polling. The mean (median) length of time from start to finish for the five questions with the full method is 13.1 (12.0) minutes, while the confidence range variation is 7.6 (6.7) minutes. Further, the confidence range responses provide a statistically significant positive correlation between confidence and accuracy that can be utilized for the creation of certainty-weighted forecasts. The goal of this article is to provide validation of my method versus the best and most practical of the other possible options on information and utility, but if time/cost is an issue, there will definitely be scenarios where the confidence range is the right option.

---

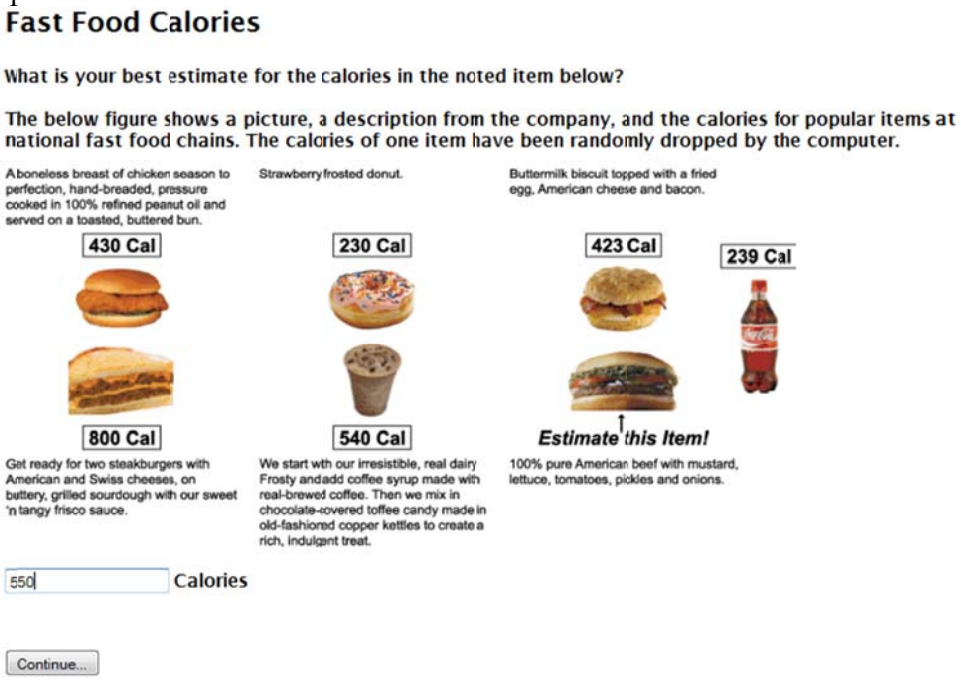
[53] There is evidence that the national popular vote prediction markets may suffer from manipulation by people motivated to gain publicity for their chosen candidate. The aggregation is over willingness to invest money, not confidence!

Turning to decision making, there is consensus in the literature on the importance of expectations in decision making. Manski (2004) demonstrates that playing a simple economic game, a subject with one of three different expectations and one of two different utility functions will make the same move (i.e., revealed behavior) in four out of six possible scenarios. He outlines many empirical examples of subjects having faulty expectations, but emphasizes the gap in the literature in understanding expectations separated from utility.

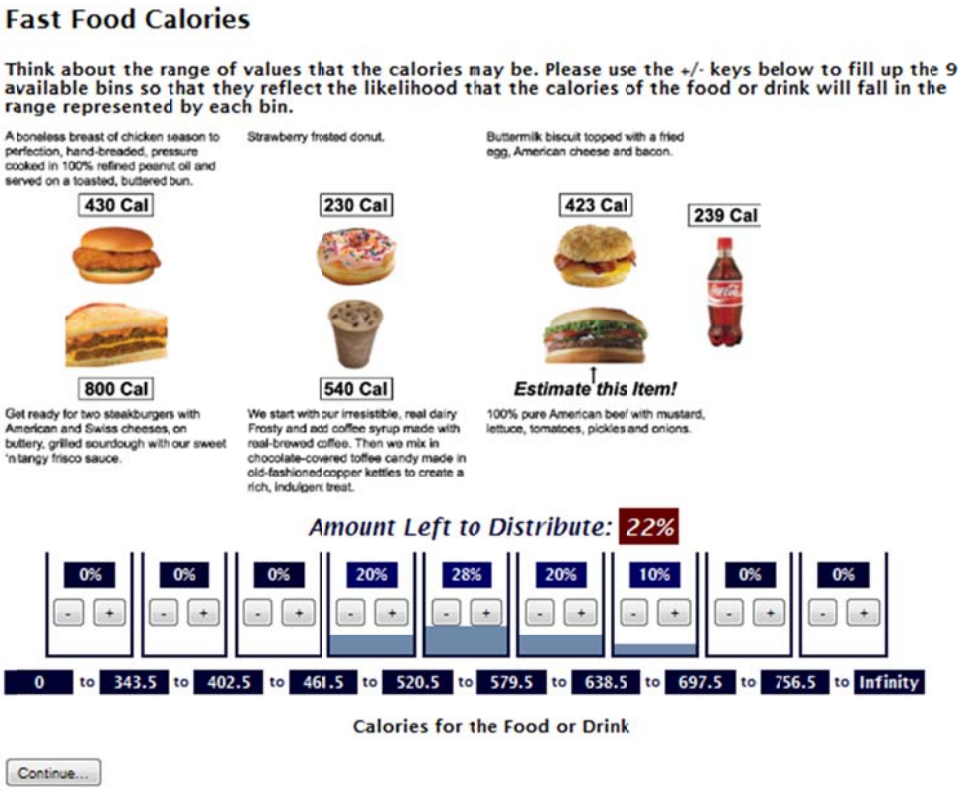
The follow-up question for the gas question hints at the usefulness of my method in decoupling expectation from utility in revealed decisions. The main question asks the respondent to imagine that she is driving down a major highway and she notes the price of gas at last few consecutive stations. She is running low and can hold out only long enough to stop at one of the next three stations; she is asked to create a probability distribution of the lowest price of gas among these next three stations. The follow-up question asks what price would induce the driver to stop at the first station she sees, rather than keep going and try one of the following two stations. The median response was at the 30% point of the probability distribution of what they expect the lowest price of the next three gas stations to be. That means that the median driver would stop where they believe that there is only a 30% chance that one of the next two stations would be less. Just 6% of respondents said they would stop at station in the 80% percentile or higher. Most importantly, there is a statistically significant positive correlation between the point-estimate expectation and price in which the driver would stop for gas. Thus, the higher the driver expects the lowest price to be, the higher price the driver will stop and pay. Further, confidence demonstrates a meaningful role in the decision making; if two drivers have the same point-estimate, the driver with the larger standard deviation (i.e., less confidence) will stop at a gas station with a higher price. Expectation matters, but so does confidence!

V. Appendix: *Sample Questions can be found on my website: www.PredictWise.com*

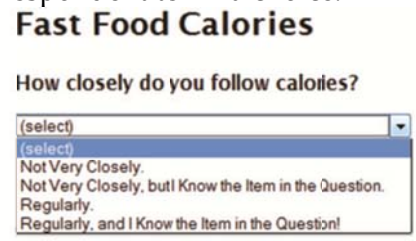
Study I & II (Calories Count): Here is a full example. It starts with the point-estimate question:



Calories Count Probability Question: Since I placed 550 calories at my estimate the probability question is centered on 550 calories:



Calories Count Follow-up Question: All drop down menus start at select, forcing the respondent to fill a choice.



Below are the other four questions used in this study:

Study I Only (Concert Ticket Prices): What is your best estimate for the lowest possible price of 2 tickets to the noted concert the day before the show? The top table highlights one randomly selected upcoming concert from the full list of StubHub's top selling concert tours. The bottom table shows the lowest possible price (including all charges) for 2 tickets, from either the box office or StubHub, whichever is lower, that could be purchased 10 days and 1 day from the concert for a random selection of concerts by performers on the same list.

Study I & II (Gas Prices): What is your best estimate for the LOWEST price of gas among the next 3 stations on the highway described in the below table? The below table shows the price of gas at the 3 previous gas stations and the question assumes that you continue down the same highway. All prices are from 8/4/2010 on a major Eastern highway. (9/16/2010 for the Study II.)

Gas Prices Follow-up Question: If you had about enough gas where you felt comfortable driving for up to 3 more stations, what price of gas would induce you to stop at the next station?

Study I Only (Movie Receipts): What is your best estimate for the 4 week gross for *MOVIE* in millions of dollars (i.e. what will *MOVIE* gross domestically through its 4th weekend of wide release)? *DESCRIPTION OF MOVIE*. Nationwide release on *DATE OF RELEASE*. The below table shows the domestic gross for the last 30 wide-release movies through their 4th weekend of release.

Movie Receipts Follow-up Question: Are you interested in seeing *MOVIE*?

Study I & II (Unemployment Rate): What is your best estimate for the August/September Unemployment Rate in the state noted below (use the slider below to show your answer)? The below table shows the Unemployment Rate in a randomly chosen state in a few relevant periods. Unemployment rates are adjusted for seasonal trends.

Unemployment Rate Follow-up Question: How familiar are you with the state in the question?

VI. References

Ariely, Dan, George Lowenstein, and Drazen Prelec. 2003. "Coherent Arbitrariness." *Quarterly Journal of Economics*, 118(1):73-105.

Artinger, Floriean, Filippos Exadaktylos, Hannes Koppel, and Lauri Sääksvuaori. 2010. "Applying Quadratic Scoring Rule Transparently in Multiple Choice Settings: A Note." Jena Economic Research Papers 2010-021.

Avery, Christopher, Judith Chevalier, and Richard Zeckhauser. 2009. "The "CAPS" Prediction System and Stock Market Returns."Working paper, Harvard University.

Barber and Odean. 2001. "Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment," *Quarterly Journal of Economics*, 116(1):261-292.

Bates J., and Grange C. 1969. "The Combination of Forecasts." *Operational Research Quarterly*, 20:451-468.

Becker, G.M., Degroth, M.H., Marashak, J. 1964. "Measuring Utility by a Single-Response Sequential Method." *Behavioral Science* 9(2):226-232.

Chen, Kay-Yut, Leslie Fine, Bernardo Huberman. 2003. "Predicting the Future." *Information Systems Frontiers*, 5(1):47-61.

Clemen, Robert. 1989. "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting*. 5:559-583.

Clemen, Robert and Robert Winkler. 1993. "Aggregating Point Estimates: A Flexible Modeling Approach." *Management Science*, 39(4):501-515.

Delavande, A. and Rohwedder, S. 2008, "Eliciting Subjective Probabilities in Internet Surveys." *Public Opinion Quarterly*, 72(5):886-891.

Dominitz, Jeff and Charles Manski.. 2006. "Measuring Pension-benefit Expectations Probabilistically." *Labour*, 20:201-236.

Engleberg, Joseph, Charles Manksi, and Jared Williams. 2009. "Comparing the Point Prediction and Subjective Probability Distributions of Professional Forecasters." *Journal of Business and Economics Statistics*, 27(1):30-41.

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.

Gurkaynak, Refet and Justin Wolfers. 2006. "Macroeconomic Derivatives: An Initial Analysis of Market-Based Macro Forecasts, Uncertainty, and Risk." CEPR Discussion Paper No. 5466.

Kahneman & Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica*, 47:263-291.

----------. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty*, 5(4):297-323.

Kaufman-Scarborough, Carol, Maureen Morrin, and Eric T. Bradlow. 2010. "Improving the Crystal Ball: Harnessing Consumer Input to Create Retail Prediction Markets." *Journal of Research in Interactive Marketing*, 4(1):30-45.

Klayman, Joshua, Jack Soll, Cladia Gonzalez-Valejo, and Sema Barlas. 1999. "Overconfidence: It Depends on How, What, and Whom You Ask." *Organizational Behavior and Human Decision Process.* 79:216-247.

Kuklinski, James, Paul Quirk, Jennifer Jerit, David Schwider, and Robert Rich. 2000. "Misinformation and the Currency of Democratic Citizenship." *The Journal of Politics*, 62(3) 790-816.

Manski, Charles. 2004. "Measuring Expectations." *Econometrica*, 72(5):1329-1376.

Pennock and Reeves. 2007. "How and When to Listen to the Crowd." http://www.overcomingbias.com/2007/02/how_and_when_to.html.

Rothschild, David. 2009. "Forecasting Elections, Comparing Prediction Markets, Polls, and Their Biases." *Public Opinion Quarterly* 73(5):895-16.

Rothschild, David and Justin Wolfers. 2011. "Forecasting Elections: Voter Intentions versus Expectations." Working paper, University of Pennsylvania, Available at: http://assets.wharton.upenn.edu/~rothscdm/RothschildExpectations.pdf.

Smith, Jeremy, and Kenneth Frank Wallis. 2009. "A Simple Explanation of the Forecast Combination Puzzle." *Oxford Bulletin of Economics and Statistics*, 71(3):3331-355.

Soll, JB and Klayman J. 2004. "Overconfidence in Interval Estimates." *Journal of Experimental Psychology Learning Memory and Cognition*, 30(20):299-314.

Sonnemans, Joep and Theo Offerman. 2001. "Is the Quadratic Scoring Rule Really Incentive Compatible?" Working paper, CREED, University of Amsterdam.

Speirs-Bridge, Andrew, Fiona Fidler, Marissa McBride, Louisa Flander, Geoff Cumming, and Mark Burgman. 2010. "Reducing Overconfidence in the Interval Judgments of Experts." *Risk Analysis*, 30(3):512-523.

Stock, James and Mark Watson. 2004. "Combination Forecasts of Output Growth in a Seven-Country Data Set." *Journal of Forecasting*, 23:405-430.

Surowiecki, James. 2004. *Wisdom of the Crowds: Why the Many are Smarter Than the Few and How Collective Wisdom Shapes Business, economics, Societies, and Nations*, Little, Brown.

Teigen, KH and M. Jorgensen. 2005. "When 90% Confidence Intervals are 50% Certain: On the Credibility of Credible Intervals." *Applied Cognitive Psychology*, 19:455-475.

Vul, Edward and Harold Pashler. 2008. "Measuring the crowd within: Probabilistic representations within individuals." *Psychological Science*, 19(7):645–647.