

# Information, incentives, and goals in election forecasts\*

Andrew Gelman<sup>†</sup>

Jessica Hullman<sup>‡</sup>

Christopher Wlezien<sup>§</sup>

8 Sep 2020

## Abstract

Presidential elections can be forecast using information from political and economic conditions, polls, and a statistical model of changes in public opinion over time. We discuss challenges in understanding, communicating, and evaluating election predictions, using as examples the Economist and Fivethirtyeight forecasts of the 2020 election.

## 1. Forecasting presidential elections

In July, 1988, Michael Dukakis, the Democratic nominee for president, was leading his Republican opponent, George H. W. Bush, by 14 points in the polls. However, there was a consensus among political scientists that the strength of the economy would likely ensure the incumbent Republicans' return to power.

During the following months, as Bush gained the lead and then won the election, there were many discussions in the news media about what he did right in his campaign—and what Dukakis did wrong—to ensure the come-from-behind victory. But in the years since, we have come to the view expressed by Abramowitz (1988) in an article predicting a narrow Bush win in that year: “The outcomes of presidential elections can be predicted with a high degree of accuracy by a simple model based on three independent variables: the incumbent president’s approval rating in the Gallup Poll, the change in real GNP during the election year, and the timing of the election.” From that perspective, the role of the general campaign is then not to persuade voters so much as to push them to where they were going anyway (Gelman and King, 1993), what Campbell (2000) calls “the predictable campaign.”

In the succeeding elections, political journalists have become increasingly aware that the election can be (approximately) forecast months ahead of time. Such forecasts are not so useful in close elections, but they have changed our views of the role of campaigning (Erikson and Wlezien, 2012, Jacobson, 2015). It now is generally understood that campaigns largely serve to “deliver” the fundamentals of the election, keeping in mind that these do evolve some over the election timeline even as the campaign plays out. In other words, the equilibrium of an election is not fixed at some arbitrary point before (or during) the campaign.

In the meantime, polling has become cheaper, both at the national and state levels, and consumers of the news now can get daily updates on the state of the presidential race.

---

\*We thank Joshua Goldstein, Elliott Morris, Merlin Heidemanns, Dhruv Madeka, Yair Ghitza, Doug Rivers, Bob Erikson, Bob Shapiro, and Jon Baron for helpful comments and the National Science Foundation, Institute of Education Sciences, Office of Naval Research, National Institutes of Health, Sloan Foundation, and Schmidt Futures for financial support.

<sup>†</sup>Department of Statistics and Department of Political Science, Columbia University, New York.

<sup>‡</sup>Department of Computer Science & Engineering and Medill School of Journalism, Northwestern University.

<sup>§</sup>Department of Government, University of Texas at Austin.

Several news organizations now offer not just poll aggregation but probabilistic forecasts, giving the estimated chances of candidates winning individual states, the national vote, and the electoral college.

There is no agreed-upon method for combining available information into a projection for November, and forecasts coming from different groups can look much different. At the beginning of September, the Economist magazine forecast (which one of us was involved in constructing) assigned Joe Biden an 87% chance of winning the election, Fivethirtyeight.com gave him 71%, and the prediction market Betfair called it a tossup. How do these forecasts work? What are the goals of election forecasters, and what incentives do they have to get things right? We discuss these issues in the context of election forecasting, recognizing the connection to more general concerns of communication about uncertainty.

In this article, we use as examples our Economist forecast, which uses a statistical model that elaborates on those of Linzer (2013) and earlier work in political science, and the Fivethirtyeight forecast, which is notable both as a trailblazing nonacademic effort and as a part of a larger project of high-quality data analysis and communication of predictive uncertainty in sports and science as well as politics. As we discuss below, the Economist and Fivethirtyeight produce similar forecasts despite using much different statistical approaches. This should be no great surprise given that the two organizations are using roughly the same information (state and national polls, previous election results, and current political and economic conditions) and both procedures have been calibrated to past elections. When we do discuss differences between the two forecasts, this is not to disparage either approach but rather to point to directions of potential improvements in modeling and communication.

We are speaking here only of general elections for president, which are relatively predictable due to having highly visible campaigns lasting several months, typically with only two major candidates (Gelman, 2011). Prediction can be much more difficult for primaries and local offices, and we do not consider these here, and there is also substantial variation across countries (see, for example, Jennings and Wlezien, 2016).

### **1.1. Forecasting elections from political and economic fundamentals**

There is a large literature in political science and economics about factors that predict election outcomes; notable contributions include Fair (1978), Fiorina (1981), Rosenstone (1983), Holbrook (1991), Campbell (1992), Lewis-Beck and Rice (1992), Wlezien and Erikson (1996), and Hibbs (2000). That research finds that the incumbent party candidate typically does better in times of strong economic growth, high presidential approval ratings, and when the party has not seeking a third consecutive term. This last reflects a “cost of ruling” effect, which has been shown to impact elections around the world (Paldam, 1986, Cuzan, 2015).

Although these referendum judgments are important for presidential elections, the electoral choice is too. For instance, candidates gain votes by moving toward the median voter (Erikson, MacKuen, and Stimson, 2002). And voters also matter. Perhaps most notably at present, partisanship can influence the impact of economics and other short-term forces (Kayser and Wlezien, 2011, Abramowitz, 2012). The various fundamentals of an election increasingly become reflected in—and evident from—the polls (Erikson and Wlezien, 2012).

These general ideas are hardly new; for example, a prominent sports oddsmaker described how he handicapped presidential elections in 1948 and 1972 based on the relative

strengths and weaknesses of the candidates (Snyder, 1975). But one value of a formal academic approach to forecasting is that it can better allow integration of information from multiple sources, by systematically using the sources of information that appear to have been predictive in the past.

At a state level, the relative positions of the states usually don't change much from one election to the next, with the major exceptions in recent decades being some large swings in the south during the period from the 1950s through the 1980s as that region swung toward the Republicans.

With the increase in political polarization in recent decades, there is reason to believe that elections should be both more and less predictable than in the past: more predictable in the sense that voters are less subject to election-specific influences as they will just vote their party anyway, and less predictable in that elections should be closer to evenly balanced contests. To put it another way, a given uncertainty in the predicted *vote share* for the two parties corresponds to a much greater uncertainty in the election *outcome* if the forecast vote share is 50/50 than if it is 55/45.

## 1.2. Pre-election surveys and poll aggregation

Political campaigns have, we assume, canvassed potential voters for as long as there have been elections, and the Gallup poll in the 1930s propagated general awareness that it is possible to learn about national public opinion from surveys. Indeed, even the much-maligned Literary Digest poll of 1936 would not have performed so badly had it been adjusted for demographics in the manner of modern polling (Lohr and Brick, 2017). The ubiquity of polling has changed the relationship between government and voters, which George Gallup and others have argued is good for democracy (Igo, 2006), while others have offered more sinister visions of voter manipulation (Burdick, 1964).

In any case, polling has moved from in-person interviews to telephone calls to autodialing to the internet, each time reducing costs, and now we are overwhelmed with state and national polls during every election season, with an expectation of a new sounding of public opinion within days of every major news event. The results produced by different survey organizations differ in a variety of ways, what sometimes are referred to as “house effects” (Erikson and Wlezien, 1999, Pasek, 2015).

With the proliferation of polls have come aggregators such as Real Clear Politics and the Huffington Post, which report the latest polls along with smoothed averages for national and state races. Polls thus supply ever more raw material for pundits, but this is happening in a politically polarized environment in which campaign polls are more stable than ever before, and even the relatively small swings that do appear can largely be attributed to differential nonresponse (Gelman, Goel, et al., 2016).

Surveys are not perfect, and a recent study of presidential, senatorial, and gubernatorial races found that state polls were off from the actual elections by about twice the stated margin of error (Shirani-Mehr et al., 2018). Most notoriously, the polls in some midwestern states overestimated Hillary Clinton's support by several percentage points during the 2016 campaign, an error that has been attributed in part to oversampling of high-education voters and a failure to adjust for this sampling problem (Gelman and Azari, 2017, Kennedy, et al., 2018). Pollsters are now reminded to make this particular adjustment (and analysts are

reminded to discount polls that do not do so), but it is always difficult to anticipate the *next* polling failure. There are also concerns about bias in partisan survey organizations, “herding” by pollsters who can adjust away discordant results, and the opposite concern of pollsters who get attention from counterintuitive claims. All these issues add challenges to poll aggregation. For a useful summary of research on pooling the polls when predicting elections, see Pasek (2015).

A single survey yields an estimate and standard error which is often interpreted as a probabilistic snapshot or forecast of public opinion: for example, an estimate of  $53\% \pm 2\%$  would correspond to an approximate 95% predictive interval of (49%, 57%) for a candidate’s support in the population. This Bayesian interpretation of a classical confidence interval is only correct in the context of a (generally inappropriate) uniform prior. With poll aggregation, however, there is an implicit or explicit time series model which, in effect, serves as a prior for the analysis of any given poll. Thus, poll aggregation should be able to produce a probabilistic “nowcast” of current vote preferences and give a sense of the uncertainty in opinion at any given time.

### 1.3. Putting together an electoral college forecast

The following information can be combined to forecast a presidential election:

- A fundamentals-based forecast of the national election,
- The relative positions of the states in previous elections, along with a model for how these might change,
- National polls,
- State polls,
- Models for sampling and nonsampling error in the polls,
- A model for state and national opinion changes during the campaign.

We argue that all these sources of information are necessary, and if any are not included, the forecaster is implicitly making assumptions about the missing pieces. One underlying model is that the polls represent mean reversion rather than a random walk (Kaplan, Park, and Gelman, 2012), but the level to which there is “reversion” is itself unknown, more precisely a reversion to slightly changing fundamentals (Erikson and Wlezien, 2012).

In practice, the key aspects of a presidential election forecast are model the vote share in each state, including both state and national polls (which in turn requires some model of underlying opinion; see Lock and Gelman, 2010, and Linzer, 2013), to model or otherwise account for nonsampling error and polling biases, and to appropriately capture correlation of uncertainties among states. This last factor is important, as our ultimate goal is an electoral college prediction. The steps of our own model for the Economist are described in Morris (2020a), but these principles should apply to any polling-based forecasting procedure.

At this point one might wonder whether a simpler approach could work, simply predicting the winner of the national election directly, or estimating the winner in each state, without going through the intermediate steps of modeling vote share. Such a “reduced form”

approach has the advantage of reducing the burden of statistical modeling but at the prohibitive cost of throwing away too much information. Consider, for example, the “13 keys to the presidency” that purportedly predicted every election for several decades (Lichtman, 1996). The trouble is that landslides such as 1964, 1972, and 1984 are so easy to predict that they supply almost no information relevant to training a model, while tie elections such as 1960, 1968, and 2000 are so close that a model should get no more credit for predicting the winner than it would for predicting a coin flip. In contrast, a forecast of vote share gets potentially valuable information from all elections (Gelman, 1993). Another advantage of predicting state-by-state vote share, as we shall discuss, is that this provides us with more opportunities for checking and understanding a national election forecast.

## 2. Communicating and diagnosing problems with probabilistic election forecasts

### 2.1. Win probabilities

There is a persistent confusion between forecast vote share and win probabilities. A vote share of 60% is a landslide win, but a win probability of 60% corresponds to an essentially tied election. For example, as of September 1, the Economist model was forecasting a 54% share of the two-party vote for Biden and an 87% chance of winning.

How many decimal places does it make sense to report the win probability? We work this out using the following simplifying assumptions: (1) each candidate’s share of the national two-vote is forecast with a normal distribution, and (2) as a result of imbalances in the electoral college, Biden wins the election if and only if he wins at least 51.7% of the two-party vote. Both of these are approximations, but generalizing to non-normal distributions and aggregating statewide forecasts will not really affect our main point here.

Given the above assumptions, suppose the forecast of Biden’s national vote share is 54% with a standard deviation of 2%. Then the probability that Biden wins can be calculated using the normal cumulative distribution function:  $\Phi((0.54 - 0.517)/0.02) = 0.875$ .

Now suppose that our popular vote forecast is off by half of a percentage point. Given all our uncertainties, it would seem pretty ridiculous to claim we could forecast to that precision anyway, right? If we bump Biden’s predicted two-party vote down to 53.5%, his win probability drops to  $\Phi((0.535 - 0.517)/0.02) = 0.816$ .

Thus, a shift of 0.5% in Biden’s expected vote share corresponds to a change of 6 percentage points in his probability of winning. Conversely, a change in 1% of win probability corresponds to a 0.1% percentage point share of the two-party vote. There is no conceivable way to pin down public opinion to a one-tenth of a percentage point, which suggests that, not only is it meaningless to report win probabilities to the nearest tenth of a percentage point, it’s not even informative to present that last digit of the percentage.

On the other hand, if we round to the nearest 10 percentage points so that 87% is reported as 90%, this creates other difficulties at the high end of the range—we would not want to round 96% to 100%—and also there will be sudden jumps when the probability moves from 90% to 80%, say. For the 2020 election, both the Economist and Fivethirtyeight compromised and rounded to the nearest percentage point but then summarized these numbers in ways intended to convey uncertainty and not lead to overreaction to small, meaningless changes.

One can also explore how the win probability depends on the uncertainty in the vote. Again continuing the above example, suppose we increase the standard deviation of the national vote from 2 to 3 percentage points. This decreases the win probability from 0.875 to  $\Phi((0.54 - 0.517)/0.03) = 0.77$ .

## 2.2. Visualizing uncertainty

There is a literature on communicating probability statements (for example, Gigerenzer and Hoffrage, 1995, Spiegelhalter, Pearson, and Short, 2011) but it remains a challenge to find effective ways to express election forecasts so they will be informative to political junkies without being misinterpreted by laypeople. Human perceptions of visualized uncertainty are often noisy and prone to heuristics, especially when readers are interested mainly in a high-level view of what the forecast says. Heterogeneity and the difficulty of defining what a reader *should* do given a probabilistic forecast make it hard to evaluate graphical displays of probabilistic predictions (Hullman et al., 2018).

In the context of Fivethirtyeight’s election forecast, Wiederkehr (2020) writes:

Our impression was that people who read a lot of our coverage in the lead-up to 2016 and spent a good amount of time with our forecast thought we gave a pretty accurate picture of the election . . . People who were looking only at our top-line forecast numbers, on the other hand, thought we bungled it. Given the brouhaha after the 2016 election, we knew we had to thoughtfully approach how we delivered the forecast. When readers came looking to see who was favored to win the election, we needed to make sure that information lived in a well-designed structure that helped people understand where those numbers are coming from and what circumstances were affecting them.

One message from the psychology literature is that it is better to express probabilities as natural frequencies to provide a more concrete impression of probability. Natural frequencies work well for examples such as disease risk (“Out of 10,000 people tested, 600 will test positive, out of whom 150 will actually have the disease”).

A frequency framing becomes more abstract when applied to a single election. Formulations such as “if this election were held 100 times” or “in 10,000 simulations of this election” are not quite as natural. Still, frequency framing may better emphasize lower probability events that readers are tempted to ignore with probability statements. When faced with a probability, it can be easier to round up (or down) than to form a clear conception of what a 70% chance means. We won’t have more than one Biden versus Trump election to test a model’s predictions on, but we can imagine applying predictions to a series of elections.

Research on frequency-framed visualizations such as quantile dotplots (Kay et al., 2016) and animated hypothetical outcomes (Hullman, Resnick, and Adar, 2015) suggests that in certain decisions contexts where they are incentivized by money, laypeople can make more accurate probability judgments and better decisions with them than with density plots, error bars, and other standard statistical graphics for conveying uncertainty; see also Fernandes et al. (2018) and Kale et al. (2018). At the same time, studies suggest that when confronted with visualizations of estimates with uncertainty, many people use heuristics, like looking

at the visual distance between mean estimates and using it to judge the reliability of a difference (Hullman, Resnick, and Adar, 2015, Kale, Kay, and Hullman, 2020). People may apply heuristics like judging visual distances which are flawed cues for effect size even when given an “optimal” uncertainty visualization for their task such as animated draws (Kale, Kay, and Hullman, 2020). So even when an uncertainty visualization makes uncertainty more concrete, it will also matter how forecasters explain it to readers and how much attention readers spend on it.

Figure 1 shows uncertainty visualizations from recent election campaigns that range from probabilistic animation to more conventional shaded error envelopes. The New York Times needle was effective at conveying uncertainty in a way that was visceral and hard to ignore, but many readers expressed disapproval and even anger at its use. While it wasn’t clear to many readers what exactly drove each movement of the needle, a bigger contributor to the disapproval may have been that the needle was very different from the standard presentations of forecasts that had been used up until election night. Readers who had relied on simple heuristics to discount uncertainty shown in static plots were suddenly required to contend with uncertainty, at a time when they were already anxious.

For illustrating the history of predictions, the New York Times time series plot is helpful. The Fivethirtyeight time series plot is clear and simple, but, as noted in Section 2.1, it presented probabilities to an inappropriately high precision given the uncertainties in the inputs to the model. In addition, readers who focus on a plot of win probability may fail to understand how this maps to vote share.

Fivethirtyeight’s dot distribution displays uncertainty in a way that looks like natural frequencies. This makes it likely to be a more effective top-level display than a single probability that a reader might simply round up or down, but the forecaster has to take care in explaining what each dot represents. The Economist plot of forecast vote has the appealing feature of being able to include the poll data and the model predictions on the same scale, but it does not map directly onto win probability.

As in presenting model predictions in general, it is good to have multiple visualizations to capture different aspects of data and the predictive distribution as they change over time. Multiple plots showing components of a forecast can implicitly convey information about the model and its assumptions. As an extreme example, a collage of different state-level maps from Fivethirtyeight’s model helped us get insight into the model’s assumptions despite the fact that the model’s code isn’t public. While we expect few readers would do this, multivariate displays comprised of multiple graphics, used over time, can help a reader get more insight into the sources of information that go into a complex model.

### **2.3. Other ways to communicate uncertainty**

It’s difficult to present an election forecast without some narrative and text expression of the results. But effectively communicating uncertainty in text might be even harder than visualizing probability. Research has found that the probability ranges people assign to different text descriptions of probability such as “probable,” “nearly certain,” and so forth, vary considerably across people (Wallsten, Budescu, and Rappaport, 1986, Budescu, Weinberg, and Wallsten, 1988).

For uncertainty that can’t be quantified because it involves unknowns like how credible

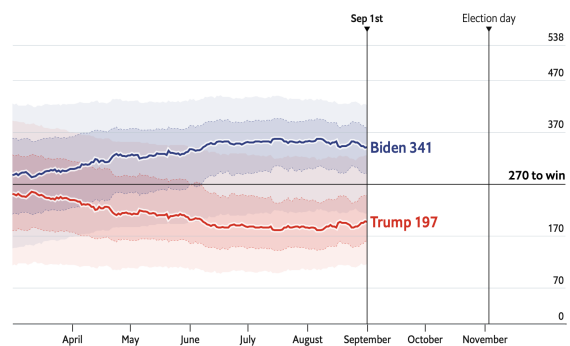
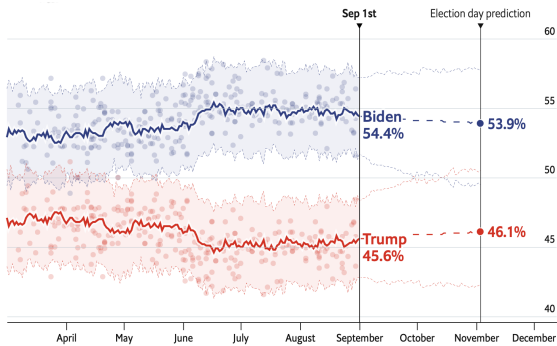
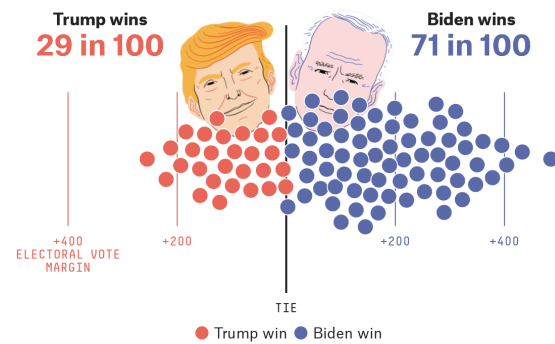
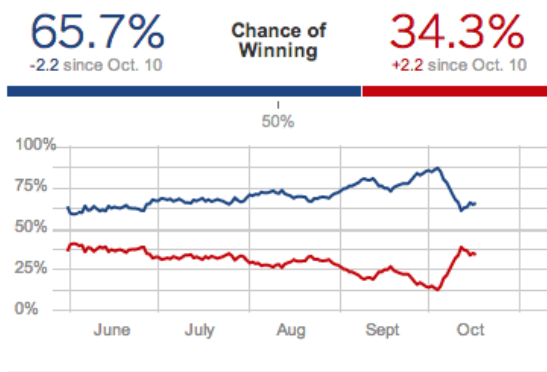
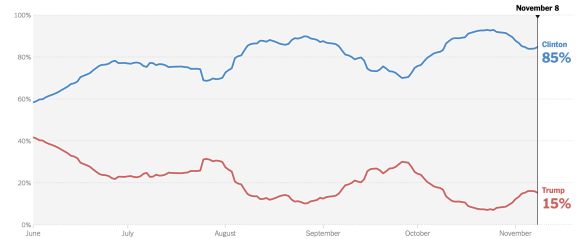
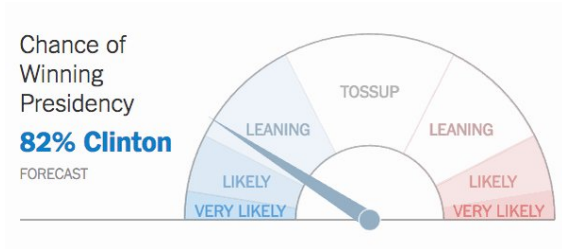


Figure 1: *Some displays of uncertainty in presidential election forecasts* Top row: *election needle and probability time series plot from the New York Times in 2016.* Center row: *time series of probabilities from Fivethirtyeight in 2012 and their dot distribution in 2020.* Bottom row: *time series of electoral and popular vote projections from the Economist in 2020.* No single visualization captures all aspects of uncertainty, and each of these displays is open to misinterpretation.



an assumption is, qualitative text expressions like “there is some uncertainty around these results due to X” may help. Some research suggests that readers may take these qualitative statements more seriously than they do quantitative cues (van der Bles, Freeman, and Spiegelhalter, 2019). Fivethirtyeight’s 2020 forecast introduces “Fivey Fox,” a bespectacled, headphones-wearing, sign-holding cartoon in the page’s margins who delivers advice directly to readers. In addition to providing guidance on reading charts and pointing to further information on the forecast, Fivey also seems intended to remind readers of the potential for very low probability events that run counter to the forecast’s overall trend, for example reminding readers that “some of the bars represent really weird outcomes, but you never know!” as they examine a plot showing many possible outcomes produced by the forecast. The problem is that how strongly these statements should be worded and how effective they are is difficult to assess, because there is no normative interpretation to be had. More useful narrative accompaniments to forecasts would include some mention of why there are unknowns that result in uncertainty. This is not to say that tips such as those of Fivey Fox are a bad idea, just that, as with other aspects of communication, their effectiveness is hard to judge and so we are relying on intuition as much as anything else in setting them up and deploying them.

Communicating uncertainty is not just about recognizing its existence; it is also about placing that uncertainty within a larger web of conditional probability statements. In the election context, these could relate to shifts in the polls or to unexpected changes in underlying economic and political conditions.

## 2.4. State and national predictions

A national vote prediction is actually a forecast of the candidates’ vote shares in the 50 states; thus, we are talking about forecasting a vector of length 50 (actually 51 including the District of Columbia, or a few more including congressional districts in Maine and Nebraska). This has implications for modeling, as we need to account for correlations in the uncertainty distribution among states, both locally (if a candidate does better than expected in North Dakota, he or she is likely to do better in South Dakota as well) and nationally (if a candidate does better than expected in any state, then on average we would expect him or her to do somewhat better all over the country), and it also has implications for understanding a fitted model.

For example, the Fivethirtyeight site gives a 95% predictive interval of (42%, 60%) for Biden’s share of the two-party vote in Florida, and it also predicts that Trump, in the unlikely event that he wins California, has a 30% chance of losing in the electoral college. Neither of these predictions seem plausible to us. The Florida interval seems too wide, given that Biden is currently at 52% in the polls there and at 54% in the national polls and in our fundamentals-based forecast, and Florida is a swing state. Other fundamentals-based forecasts put the election at closer to 50/50, but even there we don’t see how one could plausibly get to a Trump landslide in that state. The California conditional prediction, in contrast, seems too pessimistic on Trump’s chances: if the president really were to win that state, presumably this would only happen in a Republican landslide, in which case it’s hard to imagine him losing in the country as a whole.

Both the extremely wide Florida interval and the inappropriately equivocal prediction conditional on a Trump victory in California seem to reveal that the Fivethirtyeight forecast

has too low a correlation among state-level uncertainties. The model doesn't appear to account for the fact that either event—Biden receiving only 42% in Florida or Trump winning California—would in all probability represent a huge national swing.

Opinion changes in the United States tend to occur at the national level; this is the “parallel publics” identified by Page and Shapiro (1992), hence state-level swings tend to be highly correlated. Suppose you start with a forecast whose covariances across states are too low, in the sense of not fully reflecting the underlying correlations of opinion changes across states, and you want this model to have a reasonable uncertainty at the national level. To get this, you need to make the uncertainties within each state too wide, to account for the variance reduction that arises from averaging over the 50 states. Thus, implausible state-level predictions may be artifacts of too-low correlations, along with the forecasters' desire to get an appropriately wide national forecast. Low correlations can also arise if you start with a model with high correlations and then add independent state errors with a long-tailed distribution.

One reason we are so attuned to this is that a few weeks after we released our first model of the election cycle for the Economist, we were disturbed at the narrowness of some of its national predictions. In particular, at one point the model had Biden with a 99% chance of winning the national vote. Biden was clearly in the lead; at the same time, we thought that 99% was too high a probability. Seeing this implausible predictive interval motivated us to refactor our model, and we found some bugs in our code and some other places where the model could be improved—including an increase in between-state correlations, which increased uncertainty of national aggregates. The changes in our model did not have huge effects—not surprisingly given that we had tested our earlier model on 2008, 2012, and 2016—but the revision did lower Biden's estimated probability of winning the popular vote to 98%. This was still a high value, but it was consistent with the polling and what we'd seen of variation in the polls during the campaign.

The point of this discussion is not to say that the Fivethirtyeight forecast is “wrong” and that the Economist model is “right”—they are two different procedures, each with their own strengths and weaknesses—but rather that, in either case, we can interrogate a model's predictions to better understand its assumptions and relate it to other available information or beliefs.

## **2.5. Replacement candidates, vote-counting disputes, and other possibilities not included in the forecasting model**

One challenge when interpreting these forecasts is that they do not represent all possible outcomes. The 2020 election does not appear to feature any serious third-party challenges, but all the forecasts we have discussed are framed as Biden vs. Trump, so if either candidate dies or is incapacitated or is otherwise removed from the ballot before the election, it's not quite clear how to interpret the models' probabilities. We could start by just taking the probabilities to represent the Democrat vs. the Republican, and this probably would not be so far off, but a forecast will not account for that uncertainty ahead of time unless it has been explicitly included in the model. This should not be much of a concern when considering 50% intervals, but when we start talking about 95% intervals, we need to be careful about what is being conditioned on, especially when forecasts are being prepared many months

before the election.

Another concern that has arisen has been raised for the 2020 election is that people may have difficulty voting and that many votes may be lost or ruled invalid. It is not our purpose here to examine or address such claims; rather, we note that vote suppression and spoiled ballots could interfere with forecasts.

When talking about the election, we should distinguish between two measures of voting behavior: (1) *vote intentions*, the total number of votes for each candidate, if everyone who wants to vote gets to vote and if all these votes are counted; and (2) the *official vote count*, whatever that is, after some people decide not to vote because the usual polling places are closed and the new polling places are too crowded, or because they planned to vote absentee but their ballots arrived too late (as happened to one of us on primary day this year), or because they followed all the rules and voted absentee but then the post office didn't postmark their votes, or because their ballot is ruled invalid for some reason.

Both these ways of summing up—vote intentions and the official vote count—matter for our modeling, as complications owing to the latter are difficult to anticipate at this point. They are important for the country itself; indeed, if they differ by enough, we could have a constitutional crisis.

The poll-aggregation and forecasting methods we have discussed are entirely forecasts of vote intentions. Polls measure vote intentions, and any validation of forecasting procedures is based on past elections, where there have certainly been some gaps between vote intentions and the official vote count (notably Florida in 2000; see Mebane, 2004), but nothing like what it would take to get a candidate's vote share in a state from, say, 47% down to 42%. There have been efforts to model the possible effects of vote suppression in the upcoming election (see, for example, Morris, 2020b)—but we should be clear that this is separate from, or in addition to, poll aggregation and fundamentals-based forecasts calibrated on past elections.

### 3. Calibration and incentives

Election forecasting seems to be an exception to the usual rule of de-emphasizing uncertainty in data-driven reporting aimed at the public, such as media and government reporting. Forecasters also appear to be devoting more effort to better expressing uncertainty over time. Wiederkehr (2020) discusses choices made in displaying predictions for 2020, in response to criticisms of the ways in which forecasts had been presented in the previous election.

This acknowledgment that graphics that convey too much precision are risky may mean that forecasters are incentivized to express wide intervals, perceiving the loss from the interval not including the ultimate outcome to be greater than the gain from providing a narrow, precise interval. We have also heard of news editors not wanting to “call the race” before the election happens, regardless of what their predictive model says. Compared to other data reporting, a forecast may be more obvious to readers as a statement that a paper is making, so the uncertainty also has to be obvious, despite readers' tendencies to try to ignore it. At the same time, reasons to underreport uncertainty are pervasive in data reporting for broad audiences (Manski, 2019), the potential for comparisons between forecasters may shift perceived responsibility, and the public may bring expectations that news outlets continually provide new information. These factors combine to make forecasters' incentives complex.

### 3.1. The difficulty of calibration

Political forecasting poses particular challenges in evaluation. Consider that 95% intervals are the standard in statistics and social science, but we would expect a 1-in-20 event only once in 80 years of presidential elections. Even if we are willing to backtest a forecasting model on 10 past elections, this will not be nearly enough information to evaluate 95% intervals. Some leverage can be gained by looking at state-by-state forecasts, but state errors can be correlated, so these 10 national elections would not represent 500 independent data points. This is not to say that calibration is a bad idea, just that it must be undertaken carefully, and 95% intervals will necessarily be strongly model dependent.

Boice and Wezerek (2019) present a graph assessing calibration of forecasts from Fivethirtyeight based on hundreds of thousands of election predictions, but ultimately these are based on a much smaller number of events used to measure the calibration, and these events are themselves occurring in only a few election years. For a simple example, suppose we had data on 10 independent events, each forecast with probability 0.7. Then we would expect to see a 0.7 success rate, but with a standard error of  $\sqrt{0.7 \cdot 0.3/10} = 0.14$ , so any success rate between, say, 0.5 and 0.9 would be consistent with calibration. It would be possible here to diagnose only extreme cases of miscalibration.

In addition, anything we might say about over or underconfidence of forecasts is inherently speculative, as we do not typically have enough information to judge whether a political forecasting method is uncalibrated—or, to be precise, to get a sense of under what conditions a forecast will be over or underconfident.

### 3.2. Incentives for overconfidence

Less than a month before the 2016 election, cartoonist Scott Adams wrote, “I put Trump’s odds of winning in a landslide back to 98%,” a prediction that was evidently falsified—it would be hard to call Trump’s victory, based on a minority of the votes, as a “landslide”—while, from a different corner of the political grid, neuroscientist Sam Wang gave Hillary Clinton a 98% chance of winning in the electoral college, another highly confident prediction that did not come to pass (Adams, 2016, Wang, 2016). These failures did not remove either of these pundits from the public eye. As we wrote in our post-election retrospective (Gelman and Azari, 2017):

There’s a theory that academics such as ourselves are petrified of making a mistake, hence we are overcautious in our predictions; in contrast, the media (traditional news media and modern social media) reward boldness and are forgiving of failure. This theory is supported by the experiences of Sam Wang (who showed up in the New York Times explaining the polls after the election he’d so completely biffed) and Scott Adams (who triumphantly reported that his Twitter following had reached 100,000).

There are other motivations for overconfidence. The typical consumer of an election forecast just wants to know who is going to win; thus there is a motivation for the producer of a forecast to fulfill that demand which is implicit in the conversation, in the sense of Grice (1975). And, even without such any direct motivation for overconfidence, it is difficult for

people to fully express their uncertainty when making probabilistic predictions (Alpert and Raiffa, 1982, Erev, Wallsten, and Budescu, 1994).

Another way to look at overconfidence is to consider the extreme case of just reporting point forecasts without any uncertainty at all. Rationales for reporting point estimates without uncertainty include fearing that uncertainty information will imply unwarranted precision in estimates (Fischhoff, 2012); not feeling that there are good methods to communicate uncertainty (Hullman, 2019); thinking that the presence of uncertainty is common knowledge (Fischhoff, 2012); thinking that non-expert audiences will not understand the uncertainty information and resort to “as-if optimization” that treats probabilistic estimates as deterministic regardless (Fischhoff, 2012, Manski, 2019); thinking that not presenting uncertainty will simplify decision making and avoid overwhelming readers (Hullman, 2019, Manski, 2019); thinking that not presenting uncertainty will make it easier for people to coordinate beliefs (Manski, 2019); and thinking that presenting it will make the message seem less credible (Fischhoff, 2012, Manski, 2019, Hullman, 2019).

There also are direct motivations for forecasters to minimize uncertainty. If calibrated intervals are too hard to construct, it can be easier to express uncertainty qualitatively than to get a good quantitative estimate of it. In addition, expressing high uncertainty violates a communication norm and can cause readers to distrust the forecaster (Hullman, 2019, Manski, 2018). This is sometimes called the auto mechanic’s incentive: if you are a mechanic and someone brings you a car, it is best for you to confidently diagnose the problem and suggest a remedy, even if you are unsure. Even if your diagnosis turns out to be wrong, you will make some money; conversely, if you honestly tell the customer you don’t know what is wrong with the car, you will likely lose this person’s business to another, less scrupulous, mechanic.

Election forecasters are in a different position than auto mechanics, in part because of the vivid memory of polling errors such as 1948 and 2016 and in part because there is a tradition of surveys reporting margins of error. Still, there is room in the ecosystem for bold forecasters such as Lichtman (1996), who gets a respectful hearing in the news media every four years (Stevenson, 2016, Raza and Knight, 2020) with his “surefire guide to predicting the next president.”

### **3.3. Incentives for underconfidence**

One incentive to make prediction intervals wider, and to keep predictive probabilities away from 0 and 1, is an asymmetric loss function. A prediction that is bold and wrong can damage our reputation more than we would gain from one that is bold and correct. To put it another way: suppose we were only to report 50% intervals. Outcomes that fall within the interval will look from the outside like “wins” or successful predictions; observations that fall outside look like failures. From that perspective there is a clear motivation to make 50% intervals that are, say, 70% likely to cover the truth, as this will be expected to supply a steady stream of wins (without the intervals being so wide as to appear useless).

In 1992, we constructed a hierarchical Bayesian model to forecast presidential elections, not using polls but only using state and national level economic predictors as well as some candidate-level information, with national, regional, and state-level error terms. Our goal was not to provide real-time forecasts but just to demonstrate the predictability of elections;

nonetheless, just for fun we used our probabilistic forecast to provide a predictive distribution for the electoral college along with various calculations such as the probability of an electoral college tie and the probability that a vote in any given state would be decisive. One reason we did not repeat this exercise in subsequent elections is that we decided it could be dangerous to be in the forecasting business: one bad-luck election could make us look like fools. It is easier to work in this space now because there are many players, so any given forecaster is less exposed; also, once there is poll aggregation, forecasting is a logical next step.

Regarding predictions for 2020, the creator of the Fivethirtyeight forecast writes, “we think it’s appropriate to make fairly conservative choices \*especially\* when it comes to the tails of your distributions. Historically this has led 538 to well-calibrated forecasts (our 20% really mean 20%)” (Silver, 2020b). But conservative prediction corresponds can produce a too-wide interval, one that plays it safe by including extra uncertainty. In other words, conservative forecasts should lead to underconfidence: intervals whose coverage is greater than advertised.

And, indeed, according to the calibration plot shown by Boice and Wezerek (2019) of Fivethirtyeight’s political forecasts, in this domain 20% for them really means 14%, and 80% really means 88%. This track record from previous elections is consistent with Silver’s goal of conservatism, which leads to underconfidence. Underconfident probability assessments are a rational way to hedge against the reputational loss of having the outcome fall outside a forecast interval, and arguably this cost is a concern in political predictions more than in sports, as sports bettors are generally comfortable with probabilities and odds. Fivethirtyeight’s probabilistic forecasts for sporting events do appear to be calibrated (Boice and Wezerek, 2019).

Speaking generally, some rationales for unduly wide intervals—underconfident or conservative forecasts—are that they can motivate receivers of the forecast to diversify their behavior more, and they can allow forecasters to avoid the blame that arises when they predict a high-probability win for a candidate and the candidate loses. This interpretation assumes that people have difficulty understanding probability and will treat high probabilities as if they are certainties. Research has shown that readers can be less likely to blame the forecaster for unexpected events if uncertainty in the forecast has been made obvious (Joslyn and LeClerc, 2012).

### 3.4. Comparing different forecasts

Incentives could get complicated if forecasters expect “dueling certitudes” (Manski, 2011), cases where multiple forecasters are going head to head. For example, suppose a forecaster knows that other forecasters will likely be presenting estimates that will differ from hers at least partially. This could shift some of the perceived responsibility for getting the level of uncertainty calibrated to the group of forecasters. Maybe in such cases each forecaster is incentivized to have a narrower interval since the perceived payoff might be bigger if they appear to readers to better predict the outcome with a precise forecast than some competitor could. Or a forecaster might think about the scoring rule from the perspective of the reader who will have access to multiple forecasts, and try to make their model counterbalance others that they believe are too extreme.

When comparing our Economist forecast to Fivethirtyeight’s, one thing we noticed was

that, although the betting probabilities were much different—87% chance of a Biden win from our model, compared to 71% from theirs—the underlying vote forecasts were a lot closer than one might think. Our estimate and standard error for Biden’s two-party vote share is approximately  $54\% \pm 2\%$ ; theirs is roughly  $53\% \pm 3\%$ . These differences are real, but ultimately any choice between them will be based on some combination of trust in the data and methods used to construct each forecast, and plausibility of all the models’ predictions, as discussed in Section 2.4. There is no easy way to choose between  $54\% \pm 2\%$  and  $53\% \pm 3\%$ , both of which represent a moderate Biden lead with some uncertainty, and it should be no surprise that the two distributions are so similar, given that they are based on essentially the same information. As is often the case in statistical design and analysis, we must evaluate the method more than its product this one time.

### 3.5. Martingale property

Suppose we are forecasting some election-day outcome  $X$ , such as a candidate’s share of the popular or electoral college vote. At any time  $t$ , let  $d(t)$  be all the data available up to that time and let  $g(t) = E(X | d(t))$  be the expected value of the forecast on day  $t$ . So if we start 200 days before the election with  $g(-200)$ , then we get information the next day and obtain  $g(-199)$ , and so on until we have our election-day forecast,  $g(0)$ .

It should be possible to construct a forecast of a forecast, for example  $E(g(-100) | d(-200))$ , a prediction of the forecast at time  $-100$  based on information available at time  $-200$ . If the forecast is fully Bayesian, based on a joint distribution of  $X$  and all the data, the forecast should have the *martingale property*, which is that the expected value of an expectation is itself an expectation. That is,  $E(g(t) | d(s))$  should equal  $g(s)$  for all  $s < t$ .

To put this in an election forecasting context: there are times, such as in 1988, when the polls are in one place but we can expect them to move in a certain direction. Poll averages are not martingales: we can at times anticipate their changes. But a Bayesian forecast should be a martingale. A reasonable forecast by a well-informed political scientist in July, 1988, should already have accounted for the expected shift toward George H. W. Bush.

The martingale property also applies to probabilities, which are simply expected values of zero-one outcomes. Thus, if we define  $X = 1$  if Biden wins in the electoral college or 0 otherwise, and we define  $g(t)$  to be the forecast probability of a Biden electoral college win, based on information available at time  $t$ , then  $g(t)$  should be an unbiased predictor of  $g$  at any later time. One implication of this is that it should be unlikely for forecast probabilities to change too much during the campaign (Taleb, 2017).

Big events can still lead to big changes in the forecast: for example, a series of polls with Biden or Trump doing much better than before will translate into an inference that public opinion has shifted in that candidate’s favor. The point of the martingale property is not that this cannot happen, but that the possibility of such shifts should be anticipated in the model, to an amount corresponding to their prior probability. If large opinion shifts are allowed with high probability, then there should be a correspondingly wide uncertainty in the vote share forecast a few months before the election, which in turn will lead to win probabilities closer to 50%. Economists have pointed out how the martingale property of a Bayesian belief stream means that movement in beliefs should on average correspond to

uncertainty reduction, and that violations of this principle indicate irrational processing (Augenblick and Rabin, 2018).

The forecasts from Fivethirtyeight and the Economist are *not* fully Bayesian—the Fivethirtyeight procedure is not Bayesian at all, and the Economist forecast does not include a generative model for time changes in the predictors of the fundamentals model—that is, the prediction at time  $t$  is based on the fundamentals at time  $t$ , not on the forecasts of the values these predictors will be at election day—and thus we would not expect these predictions to satisfy the martingale property. This represents a flaw of these prediction models (along with other flaws such as data problems and the difficulty of constructing between-state covariance matrices). Similar flaws can be expected for political prediction markets; as Aldous (2013) puts it, “compared to stock markets, prediction markets are often thinly traded, suggesting that they will be less efficient and less martingale-like.”

### 3.6. Novelty and stability

A challenge when producing forecasting for a news organization is that there is a desire for new developments every day—but the election forecast can be stable for months. In any given day or week, there will be a few new polls and perhaps some new economic data, but this information should not shift the election-day prediction on average (recall the martingale property), nor in practice will one week’s data do much to change the prognosis, except in those cases where the election is on a knife edge already. Indeed, the better the forecast, the less likely it is to produce big changes during the campaign. In the past, large changes in election projections have arisen from insufficiently accounting for fundamentals (as when pundits in 1988 followed early polls and thought Dukakis had a huge lead) or from not accounting for systematic polling error (as with the apparent wide swings in 2012 and 2016 that could be explained by differential nonresponse and the state polls in 2016 that did not adjust for education; Gelman, Goel, et al., 2016, Gelman and Rothschild, 2016, Kennedy et al., 2018). As discussed, events can have big effects on the fundamentals, but such events are rare (Erikson and Wlezien, 2012).

Good forecasts thus should be stable most of the time. But from a journalistic perspective there is a push for news. One way to create news is to report daily changes in the predicted win probabilities, essentially using the forecast as a platform for punditry. But, as discussed in Section 2.1, small changes in win probabilities are essentially pure noise, with a 1% change in probability corresponding to a swing of only a tenth of a percentage point in the predicted vote share. Another way to create news is to flip this around and to report every day that, again, there is essentially no change, but this gets old fast. And the challenge of explaining that there are no real changes in the predictive distribution is that the distribution itself still is uncertain. Our 95% interval for Biden’s vote share can remain stable at around (50%, 58%) for weeks, and our 95% interval for his electoral vote total can remain steady around the interval (250, 420), but this still doesn’t tell us where the outcome will end up on election day. Stability of the forecast is *not* the same as predictability of the outcome; indeed, in some ways these two are opposed (Taleb, 2017).

We are not fans of Twitter and its 24-hour debate culture, but one advantage of this format is that it allows journalists to remain active without needing to supply any actual news. A forecaster can contribute to an ongoing discussion on social media without feeling



the need for his or her forecast to supply a continuing stream of surprises. Traditional political pundits don't seem to have yet realized this point—they continue to breathlessly report on each new poll and speculate on the polls to come—but serious forecasters, including those at Fivethirtyeight and the Economist—recognize that big news is, by its nature, rare. Rather than supplying “news,” a forecast provides a baseline of expectation that allows us to interpret the real political news as it happens.

Again, all this refers to the general election for president. Primary elections and other races can be much harder to predict and much more volatile, making forecasting a more challenging task with much more expectation of surprises.

## 4. Discussion

In the wake of the 2016 debacle, some analysts have argued that “marketing probabilistic poll-based forecasts to the general public is at best a disservice to the audience, and at worst could impact voter turnout and outcomes” (Jackson, 2020). While there surely are potential costs to forecasting, there also are benefits. First, the popularity of forecasts reflects revealed demand for such information. Second, by collecting and organizing relevant information, a forecast can help people make better decisions about their political (and economic) resources. Third, the process of building—and evaluating—forecasts can allow scholars and political observers to better understand voters and their electoral preferences, which can help us understand and interpret the results of elections.

This is not to say that creating a good forecast is easy, or that the forecaster has no responsibilities. Our discussion above implies several high level responsibilities:

- *Fundamentals.* Forecasters should be mindful of known regularities in election results. Omitting information that research indicates has predictive power makes very strong assumptions that should be explained. Regularizing public opinion polling through aggregation is one example of an approach that, if not used, makes a forecast questionable.
- *Statistical coherence.* Forecasters have a responsibility to use statistics properly, including not implying unreasonable precision, acknowledging the sensitivity of their results to assumptions, and recognizing the constraints that make it difficult to assess model calibration.

As mentioned above, neither of the forecasts under discussion are fully Bayesian, meaning that martingale properties of a Bayesian belief stream can't be expected to hold. Still, analyzing the models in terms of movement (how much a prediction varies over time) and uncertainty reduction given the net effects of information, which are equal in expectation for a Bayesian, may be informative for identifying biases that are not necessarily visible at any given time point (Augenblick and Rabin, 2018).

Responsibilities toward uncertainty communication are harder to outline. As discussed in Section 2.1, summaries such as win probabilities depend strongly on difficult-to-test assumptions, hence it is important for forecasters to air their assumptions. While opening all aspects of the model, including the code, provides the most transparency, detailed descriptions of model details can suffice for allowing discussion.

Journalists and academics alike use terms such as “horse race” and “forecast wars” in reference to election prediction, but we see forecasting as an essentially collaborative exercise. Comparative discussions of forecasts, like model comparisons in an analysis workflow, provide insight into how different assumptions about a complex process affect our ability to predict it. When the public has a chance to observe or even participate in these discussions, the benefits are greater.

In addition to thinking about what they should know, it seems relatively inarguable that forecasters have some responsibility to take into account what readers may do with a visualization or statement of a forecast’s results. That people rely on heuristics to reduce uncertainty and want simple answers is a challenge every data analyst must contend with in communicating results. In this sense we disagree with the quote that led off this section. While some people may not seem capable of interpreting probabilistic forecasts, withholding data treats this as an immutable fact. Research on uncertainty communication, however, shows that for specific contexts and tasks some representations of model results express uncertainty better than others; see also Westwood, Messing, and Lelkes (2020) and Urminsky and Shen (2019) for attempts to empirically evaluate election-specific choices about communicating predictions.

Forecasters should acknowledge the difficulties in evaluating uncertainty communication, just as they should acknowledge the difficulties in evaluating model calibration. Readers are certainly not homogeneous, heuristics can look like accurate responses, and normative interpretations often don’t exist (Hullman et al., 2018). However, these challenges don’t warrant throwing in the towel, nor pretending that communication strategies don’t matter. Instead, we think that better forecast communication might result if forecasters were to think more carefully about readers’ possible implicit reference distributions and internal decision criteria (Gelman, 2004, Hullman et al., 2018, Hullman, 2019). Designing complex cognitive models to predict decision making from election forecasts may not be realistic given the heterogeneity of forecast consumers and available resources. However, designing a forecast without any thought to how it may play into readers’ decisions seems both impractical and potentially unethical.

We argue that more attempts to prompt readers to consider model assumptions and other sources of hard-to-quantify uncertainty are helpful for producing a more literate base of forecast consumers. A skeptic might ask, if people can’t seem to understand a probability, how can we expect them to conceive of multiple models at the same time? The progression of forecast displays over time, with generally positive reception from the public (less a few misunderstood displays like the New York Times needle), suggests that laypeople can become more savvy in interpreting forecasts.

It may be instructive to investigate how consumers of election forecasts reconcile differences between forecasts or combine them to form belief distributions, so as to better understand how beliefs are formed in the forecast landscapes that characterize modern presidential elections. Combining forecasts more formally is an intriguing idea, with ample literature describing benefits of combining expert forecasts even when one forecast is clearly more refined (or in game theoretic terms, dominates others); see Clemen (1989). However, much of this literature assumes that any given expert forecast is well-calibrated, or that forecasts are Bayesian. It’s not clear that combining full election forecasting models would be equally instructive due to calibration assessment challenges (Graefe et al., 2014).

One theme of the present article is that forecasters will inevitably have their own goals and incentives. As in scientific discussions of claims, forecasters' analyses happen in a complex web of constraints and communication norms, particularly in a news context. Discussions of incentives should not be considered taboo or non-scientific, either when talking to or about election forecasters. In fact, we believe there is a need for more reflection, and research on, how incentives may shape forecasts. We are aware of some academic discussions from economists and psychologists of incentives in constructing probabilistic forecasts (Manski, 2011, 2018, Fischhoff, 2012, Baron et al., 2014). The economic and decision theoretic literature on scoring and aggregating individual predictions may also be informative.

We started this article with a story about political scientists whose models led them to distrust early polls. We end with another story, this time about broadcast journalists (MacNeil, 2019). On election night 1952, CBS used a UNIVAC computer implementing a model developed by statistician Max Woodbury to predict the winner as part of their live television forecast. Prior to closing of all the polls, the computer's prediction was that Eisenhower would collect 438 electoral votes and Stevenson 93, giving 100 to 1 odds in favor of Eisenhower.

Opinion polls, had however, shown Stevenson in the lead. CBS suggested this couldn't be right, and asked Woodbury to reexamine his algorithm. He did, and running the model again revealed a new prediction of 8 to 7 odds in favor of Eisenhower, which Walter Cronkite reported on air. Woodbury then purportedly realized he had missed a zero in re-entering the input data, and indicated to CBS that the original odds had been correct. Only when the final results came in—442 to 89 for Eisenhower—did CBS admit the cover-up to their viewers.

Reflecting on election forecasting has many lessons to teach us—about historically-demonstrated fundamentals, statistics, uncertainty communication, and incentives—but only if we are willing to listen. Fortunately, when we make public predictions using open data and code, we have many opportunities to learn.

## References

- Abramowitz, A. L. (1988). An improved model for predicting presidential election outcomes. *PS: Political Science and Politics* **21**, 843–847.
- Abramowitz, A. L. (2012). Forecasting in a polarized era: The time for a change model and the 2012 presidential election. *PS: Political Science and Politics* **45**, 618–619.
- Adams, S. (2016). The bully party. *Scott Adams Says*, 25 Oct. [www.scottadamssays.com/2016/10/25/the-bully-party](http://www.scottadamssays.com/2016/10/25/the-bully-party)
- Aldous, D. J. (2013). Using prediction market data to illustrate undergraduate probability. *American Mathematical Monthly* **120**, 583–593.
- Alpert, M., and Raiffa, H. (1982). A progress report on the training of probability assessors. In *Judgment Under Uncertainty: Heuristics and Biases*, ed. D. Kahneman, P. Slovic, and A. Tversky, 294–305. Cambridge University Press.
- Augenblick, N., and Rabin, M. (2018). Belief movement, uncertainty reduction, and rational updating. Haas School of Business, University of California, Berkeley. [faculty.haas.berkeley.edu/n Augenblick/papers/Belief%20Movement,%20Uncertainty%20Reduction,%20and%20Rational%20Updating.pdf](http://faculty.haas.berkeley.edu/n Augenblick/papers/Belief%20Movement,%20Uncertainty%20Reduction,%20and%20Rational%20Updating.pdf)

[berkeley.edu/ned/AugenblickRabin\\_MovementUncertainty.pdf](http://berkeley.edu/ned/AugenblickRabin_MovementUncertainty.pdf)

- Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., and Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis* **11**, 133–145.
- Boice, J., and Wezerek, J. (2019). How good are Fivethirtyeight forecasts? [projects.fivethirtyeight.com/checking-our-work](http://projects.fivethirtyeight.com/checking-our-work)
- Budescu, D., Weinberg, S., and Wallsten, T. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance* **14**, 281–294.
- Burdick, E. L. (1964). *The 480*. McGraw Hill.
- Campbell, J. E. (1992). Forecasting the presidential vote in the states. *American Journal of Political Science* **36**, 386–407.
- Campbell, J. E. (2000). *The American Campaign: U.S. Presidential Campaigns and the National Vote*. Texas A&M University Press.
- Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* **5**, 559–583.
- Cuzan, A. (2015). Five laws of politics. *PS: Political Science and Politics*, **48**, 415–419.
- Erev, I., Wallsten, T. S., and Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review* **101**, 519–527.
- Erikson, R. S., MacKuen, M. B., and Stimson, J. A. (2002). *The Macro Polity*. Cambridge University Press.
- Erikson, R. S., and Wlezien, C. (1999). Presidential polls as a time series: The case of 1996. *Public Opinion Quarterly* **63**, 163–177.
- Erikson, R. S., and Wlezien, C. (2012). *The Timeline of Presidential Elections*. University of Chicago Press.
- Fair, R. C. (1978). The effect of economic events on votes for president. *Review of Economics and Statistics* **60**, 159–173.
- Fernandes, M., Walls, L., Munson, S., Hullman, J., and Kay, M. (2018). Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.
- Fiorina, M. (1981). *Retrospective Voting in American National Elections*. Yale University Press.
- Fischhoff, B. (2012). Communicating uncertainty: Fulfilling the duty to inform. *Issues in Science and Technology* **28**, 63–70.
- Gelman, A. (1993). Review of *Forecasting Elections*, by M. S. Lewis-Beck and T. W. Rice. *Public Opinion Quarterly* **57**, 119–121.
- Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, **13**, 755–779.
- Gelman, A. (2009). How did white people vote? Updated maps and discussion. Statistical Modeling, Causal Inference, and Social Science, 11 May. [statmodeling.stat.columbia.edu/2009/05/11/discussion\\_and](http://statmodeling.stat.columbia.edu/2009/05/11/discussion_and)
- Gelman, A. (2011). Why are primaries hard to predict? *New York Times*, 29 Nov. [campaignstops.blogs.nytimes.com/2011/11/29/why-are-primaries-hard-to-](http://campaignstops.blogs.nytimes.com/2011/11/29/why-are-primaries-hard-to-)

## predict

- Gelman, A., and Azari, J. (2017). 19 things we learned from the 2016 election (with discussion). *Statistics and Public Policy* **4**.
- Gelman, A., Goel, S., Rivers, D., and Rothschild, D. (2016). The mythical swing voter. *Quarterly Journal of Political Science* **11**, 103–130.
- Gelman, A., and King, G. (1993). Why are American presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science* **23**, 409–451.
- Gelman, A., and Rothschild, D. (2016). Trump’s up 3! Clinton’s up 9! Why you shouldn’t be fooled by polling bounces. *Slate*, 5 Aug. [slate.com/news-and-politics/2016/08/dont-be-fooled-by-clinton-trump-polling-bounces.html](https://www.slate.com/news-and-politics/2016/08/dont-be-fooled-by-clinton-trump-polling-bounces.html)
- Gigerenzer, G., and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* **102**, 684–704.
- Graefe, A., Armstrong, J. S., Jones, R., and Cuzan, A. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting* **30**, 43–54.
- Grice, H. P. (1975). Logic and conversation. In *Syntax and Semantics*, volume 3, ed. P. Cole and J. Morgan, 41–58. Academic Press.
- Hibbs, D. (2000). Bread and peace voting in U.S. presidential elections. *Public Choice* **104**, 149–180.
- Holbrook, T. M. (1991). Presidential elections in space and time. *American Journal of Political Science* **35**, 91–109.
- Hullman, J. (2020). Why authors don’t visualize uncertainty. *IEEE Transactions on Visualization and Computer Graphics* **26**, 130–139.
- Hullman, J., Qiao, X., Correll, M., Kale, A., and Kay, M. (2018). In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics* **25**, 903–913.
- Hullman, J., Resnick, P., and Adar, E. (2015). Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLoS One* **10**, e0142444.
- Igo, S. E. (2006). “A gold mine and a tool for democracy”: George Gallup, Elmo Roper, and the business of scientific polling, 1935–1955. *History of the Behavioral Sciences* **42**, 109–134.
- Jackson, N. (2020). Poll-based election forecasts will always struggle with uncertainty. 6 Aug. [www.centerforpolitics.org/crystalball/articles/author/natalie-jackson](https://www.centerforpolitics.org/crystalball/articles/author/natalie-jackson)
- Jennings, W., and Wlezien, C. (2016). The timeline of elections: A comparative perspective. *American Journal of Political Science* **60**, 219–233.
- Joslyn, S., and LeClerc, J. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied* **18**, 126–140.
- Kale, A., Kay, M., and Hullman, J. (2020). Visual reasoning strategies and satisficing: How uncertainty visualization design impacts effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*.

- Kaplan, N., Park, D. K., and Gelman, A. (2012). Understanding persuasion and activation in presidential campaigns: The random walk and mean-reversion models. *Presidential Studies Quarterly* **42**, 843–866.
- Kay, M., Kola, T., Hullman, J., and Munson, S. (2016). When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5092–5103.
- Kayser, Mark, and Wlezien, C. (2010). Performance pressure: Patterns of partisanship and the economic vote. *European Journal of Political Research* **50**, 365–394.
- Kennedy, C., Blumenthal, M., Clement, S., Clinton, J., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Rivers, D., Saad, L., Witt, E., and Wlezien, C. (2018). An evaluation of 2016 election polls in the United States. *Public Opinion Quarterly* **82**, 1–13.
- Kos (2009). How whites really voted in 2008. Daily Kos, 26 Mar. [www.dailykos.com/storyonly/2009/3/26/713125/-How-whites-really-voted-in-2008](http://www.dailykos.com/storyonly/2009/3/26/713125/-How-whites-really-voted-in-2008)
- Lewis-Beck, M. and Rice, T. (1992). *Forecasting Presidential Elections*. Congressional Quarterly Press.
- Lichtman, A. J. (1996). *The Keys to the White House*. Madison Books.
- Linzer, D. A. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association* **108**, 124–134.
- Lock, K., and Gelman, A. (2010). Bayesian combination of state polls and election forecasts. *Political Analysis* **18**, 337–348.
- Lohr S. L., and Brick, J. M. (2017). Roosevelt predicted to win: Revisiting the 1936 Literary Digest poll. *Statistics, Politics and Policy* **8**, 65–84.
- MacNeil, J. (2019). UNIVAC predicts election results, November 4, 1952. *EDN*, 4 Nov. [www.edn.com/univac-predicts-election-results-november-4-1952](http://www.edn.com/univac-predicts-election-results-november-4-1952)
- Manski, C. F. (2011). Policy analysis with incredible certitude. *Economic Journal* **121**, F261–289.
- Manski, C. F. (2019). The lure of incredible certitude. *Economics & Philosophy* **36**, 216–245.
- Mebane, W. R. (2004). The wrong man is president! Overvotes in the 2000 presidential election in Florida. *Perspectives on Politics* **2**, 525–535.
- Morris, G. E. (2020a). How the Economist presidential forecast works. *Economist*, 5 Aug. [projects.economist.com/us-2020-forecast/president/how-this-works](http://projects.economist.com/us-2020-forecast/president/how-this-works)
- Morris, G. E. (2020b). More mail-in voting doubles the chances of recounts in close states. *Economist*, 22 Aug. [www.economist.com/united-states/2020/08/22/more-mail-in-voting-doubles-the-chances-of-recounts-in-close-states](http://www.economist.com/united-states/2020/08/22/more-mail-in-voting-doubles-the-chances-of-recounts-in-close-states)
- Page, B., and Shapiro, R. Y. (1992). *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. University of Chicago Press.
- Paldam, M. (1986). The distribution of election results and two explanations for the cost of ruling. *European Journal of Political Economy* **2**, 5–24.
- Pasek, J. (2015). Predicting elections: Considering tools to pool the polls. *Public Opinion Quarterly* **79**, 594–619.
- Raza, N., and Knight, K. (2020). He predicted Trump's win in 2016. Now he's ready to call 2020. *New York Times*, 5 Aug. [www.nytimes.com/2020/08/05/opinion/](http://www.nytimes.com/2020/08/05/opinion/)

2020-election-prediction-allan-lichtman.html

- Rosenstone, S. J. (1983). *Forecasting Presidential Elections*. Yale University Press.
- Shirani-Mehr, H., Rothschild, D., Goel, S., and Gelman, A. (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association* **113**, 607–614.
- Silver, N. (2020a). How Fivethirtyeight’s 2020 presidential forecast works—and what’s different because of COVID-19. Fivethirtyeight, 12 Aug. [fivethirtyeight.com/features/how-fivethirtyeights-2020-presidential-forecast-works-and-whats-different-because-of-covid-19/](https://fivethirtyeight.com/features/how-fivethirtyeights-2020-presidential-forecast-works-and-whats-different-because-of-covid-19/)
- Silver, N. (2020b). Twitter thread, 1 Sep. [twitter.com/NateSilver538/status/1300825871759151117](https://twitter.com/NateSilver538/status/1300825871759151117)
- Snyder, J., with Herskowitz, M., and Perkins, S. (1975). *Jimmy the Greek, by Himself*. Chicago: Playboy Press.
- Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *Science* **333**, 1393–1400.
- Stevenson, P. W. (2026). Trump is headed for a win, says professor who has predicted 30 years of presidential outcomes correctly. *Washington Post*, 23 Sep. [www.washingtonpost.com/news/the-fix/wp/2016/09/23/trump-is-headed-for-a-win-says-professor-whos-predicted-30-years-of-presidential-outcomes-correctly](https://www.washingtonpost.com/news/the-fix/wp/2016/09/23/trump-is-headed-for-a-win-says-professor-whos-predicted-30-years-of-presidential-outcomes-correctly)
- Taleb, N. N. (2017). Election predictions as martingales: An arbitrage approach. *Quantitative Finance* **18**, 1–5.
- Urminky, O., and Shen, L. (2019). High chances and close margins: How equivalent forecasts yield different beliefs. [ssrn.com/abstract=3448172](https://ssrn.com/abstract=3448172)
- van der Bles, A. M., van der Linden, S., Freeman, A. L. J., and Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences* **117**, 7672–7683.
- Wallsten, T., Budescu, D., Rapoport, A., Zwick, R., and Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General* **115**, 348–365.
- Wang, S. (2016). Why I had to eat a bug on CNN. *New York Times*, 18 Nov. [www.nytimes.com/2016/11/19/opinion/why-i-had-to-eat-a-bug-on-cnn.html](https://www.nytimes.com/2016/11/19/opinion/why-i-had-to-eat-a-bug-on-cnn.html)
- Westwood, S. J., Messing, S., and Lelkes, Y. (2020). Projecting confidence: How the probabilistic horse race confuses and demobilizes the public. *Journal of Politics* **82**.
- Wiederkehr, A. (2020). How we designed the look of our 2020 forecast. Fivethirtyeight, 13 Aug. [fivethirtyeight.com/features/how-we-designed-the-look-of-our-2020-forecast/](https://fivethirtyeight.com/features/how-we-designed-the-look-of-our-2020-forecast/)
- Wlezien, C., and Erikson, R. S. (1996). Temporal horizons and presidential election forecasts. *American Politics Research* **24**, 492–505.