Judgment and Decision Making, Vol. 15, No. 5, September 2020, pp. 863–880

Information, incentives, and goals in election forecasts

Andrew Gelman* Jessica Hullman† Christopher Wlezien‡ George Elliott Morris§

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. However, these "knowns" about how to make a good presidential election forecast come with many unknowns due to the challenges of evaluating forecast calibration and communication. We highlight how incentives may shape forecasts, and particularly forecast uncertainty, in light of calibration challenges. We illustrate these challenges in creating, communicating, and evaluating election predictions, using the Economist and Fivethirtyeight forecasts of the 2020 election as examples, and offer recommendations for forecasters and scholars.

Keywords: forecasting, elections, polls, probability

1 What we know about forecasting presidential elections

We describe key components of a presidential election forecast based on lessons learned from research and practice.

1.1 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) & Hibbs (2000). That research finds that the incumbent party candidate typically does better in times of strong economic growth, high presidential approval ratings & when the party is not seeking a third consecutive term. This latter may reflect a "cost of ruling" effect, where governing parties tend to lose vote share the longer they are in power, which has been shown to impact elections around the world (Paldam, 1986, Cuzan, 2015).

Although these referendum judgments are important for presidential elections, political ideology also matters. Candidates gain votes by moving toward the median voter (Erikson, MacKuen & Stimson, 2002), and partisanship can influence the impact of economics and other short-term forces (Kayser and Wlezien, 2011, Abramowitz, 2012). As the campaign progresses, various fundamentals of an election increasingly become reflected in — and evident from — the polls (Wlezien and Erikson, 2004, 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 data from multiple sources, by systematically using information that appear to have been predictive in the past. In addition, understanding the successes and failures of formal forecasting methods can inform theories about public opinion and voting behavior.

With the increase in political polarization in recent decades (Abramowitz, 2010, Fiorina, 2017), there is also 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. The latter can be seen from recent election outcomes themselves, both presidential and congressional. 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, as small shifts matter more in the former than the latter.

We thank Joshua Goldstein, Merlin Heidemanns, Dhruv Madeka, Yair Ghitza, Annie Liang, Doug Rivers, Bob Erikson, Bob Shapiro, Jon Baron, and the anonymous reviewers 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.

*Department of Statistics and Department of Political Science, Columbia University, New York. Email: gelman@stat.columbia.edu.

†Department of Computer Science & Engineering and Medill School of Journalism, Northwestern University.

‡Department of Government, University of Texas at Austin.

§The Economist.

1.2 Pre-election surveys and poll aggregation

Election 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 & 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 and then in many cases to the internet, following sharp declines in response rates and increases in costs of high-quality telephone polls. 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.

With the proliferation of polls have come aggregators such as Real Clear Politics, 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 every before, and even much of the relatively small swings that do appear can be attributed to differential nonresponse (Gelman, Goel, et al., 2016).

Surveys are not perfect, and a recent study of U.S. 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. More generally, the results produced by different survey organizations differ in a variety of ways, what sometimes are referred to as "house effects" (Erikson & Wlezien, 1999, Pasek, 2015), which are more relevant than ever in the modern decentralized media landscape which features polls that vary widely in design and quality. There are also concerns about "herding" by pollsters who can adjust away discordant results, along with 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 correct only 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, evolving during the campaign, as the polls become increasingly informative about the fundamentals.

1.3 State and national predictions

Political science forecasting of U.S. presidential elections has traditionally focused on the popular vote, not the electoral college result. This allows us to estimate the national forces at work, what sometimes is referred to among electoral scholars as the "swing" between elections. But national vote predictions actually are forecasts of the candidates' vote shares in the states and the District of Columbia; thus, we are talking about forecasting a vector of length 51 (plus extra jurisdictions from the congressional districts in Maine and Nebraska), and state-by-state forecasts are important unto themselves given that the electoral college actually chooses the president. This was explicitly addressed in early forecasting efforts, including those of Rosenstone (1983) and Campbell (1992), and has been on the rise in recent election cycles; see the summary in Enns and Lagodny (2020).

The national swing is revealing about what happens in the states; while vote shares vary substantially across states, swings from election to election tend to be highly correlated, an example of what Page and Shapiro (1992) call "parallel publics." At the state level, the relative positions of the states usually do not 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 shifted toward the Republicans. Hence, predicting the national vote takes us most of the way toward forecasting the electoral college — although, as we were reminded in 2016, even small percentage deviations from uniform swing can be consequential in a close election.

These correlations have clear implications for modeling, as we need to account for them in the uncertainty distribution among states: if a candidate is doing better than expected in any state, then on average we would expect him or her to do better elsewhere. There also are more local implications, for instance, if a candidate does better than expected in North Dakota, he or she is likely to do better in South Dakota as well. These correlations also are relevant when understand-

ing and evaluating a fitted model, as we discuss in Section 2.3.

1.4 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 feature any serious third-party challenges, which simplifies choice, but all the forecasts we have discussed are framed as Biden vs. Trump. If either candidate dies or is incapacitated or is otherwise removed from the ballot before the election, it is 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 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 did not 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 U.S. itself; indeed, if they differ by enough, we could have a constitutional crisis.

The poll-aggregation and forecasting methods we have discussed really are 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, 2020c) — but we should be clear that this is separate from, or in addition to, poll aggregation and fundamentals-based forecasts calibrated on past elections.

1.5 Putting together an electoral college forecast

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

- A fundamentals-based forecast of the national vote,
- 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, capturing how the relevance of different predictors changes over time.

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. State polls are relevant because of the electoral college, and national polls are relevant for capturing opinion swings, as discussed in Section 1.3. It can be helpful to think of changes in the polls during the campaign as representing mean reversion rather than a random walk (Kaplan, Park & Gelman, 2012), but the level to which there is "reversion" is itself unknown and actually can change, so that there is reversion to slightly changing fundamentals (Erikson and Wlezien, 2012).

The use of polls requires some model of underlying opinion (see Lock & Gelman, 2010 & Linzer, 2013) to represent or otherwise account for nonsampling error and polling biases, and to appropriately capture the correlation of uncertainties among states. This last factor is important, as our ultimate goal is an electoral college prediction. The steps of the Economist model are described in Morris (2020b), but these principles apply to any poll-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 information. Consider, for example, the "13 keys to the presidency" that purportedly predicted every presidential election winner for several

decades (Lichtman, 1996). The trouble with such an approach, or any other producing binary predictions, is that landslides such as 1964, 1972, and 1984 are easy to predict, and so supply almost no information relevant to training a model. 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. A forecast of vote share, by contrast, gets potentially valuable information from all elections, as it captures the full variation. Predicting state-by-state vote share allows the forecaster to incorporate even more information and also provides additional opportunities for checking and understanding a national election forecast.

1.6 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 $g(t) = \mathrm{E}(X \mid d(t))$ 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 $\mathrm{E}(g(-100) \mid d(-200))$, 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, $\mathrm{E}(g(t) \mid d(s))$ should equal $g(s)$ for all $s < t$. In non-technical terms, the martingale property says that knowledge of the past will be of no use in predicting the future.

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: its future changes should in expectation be unpredictable, which implies that the direction of anticipated future swings in the polls should be already baked into the current prediction. 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 and 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 & 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 forecasting procedures (along with other flaws such as data problems and the difficulty of constructing between-state covariance matrices). We expect that, during the early months of the campaign, a fully generative version of the Economist model would have been less confident of a Biden victory because of the added uncertainty about November economic ratings causing a wider range of fundamentals-based predictions.

2 Why evaluating presidential election forecasts is difficult

We address fundamental problems in evaluating election forecasts, stemming from core issues in assessing calibration and challenges related to how forecasts are communicated.

2.1 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 previous elections, what often are referred to as "out-of-sample" forecasts, this will not provide 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 depend on assumptions about the tail behavior of forecasts that cannot be directly checked from past data. 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.

Boice and Wezerek (2019) present a graph assessing calibration of forecasts from Fivethirtyeight based on hundreds of thousands of election predictions, but these represent predictions of presidential and congressional elections for every state on every date that forecasts were available; 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. As a result, trying to identify over- or underconfidence of forecasts is inherently speculative, as we do not typically have enough information to make detailed judgments about 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. This is not to say that reflecting on goals and incentives in election forecasting is futile; on the contrary, we think doing so can be informative for both forecast consumers and researchers, and we discuss possible incentives in Section 3.

2.2 Win probabilities

There are also tensions related to people's desire for precise win probabilities and what these probabilities mean. 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 him winning in the electoral college.

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-party 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 too strong to claim we could forecast to that precision anyway. 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 in both win probabilities and estimates of vote share.

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 $\Phi((0.54 - 0.517)/0.03) = 0.77$.

2.3 Using anomalous predictions to improve a model

Forecasters can use the uncertainty in their predictions as benchmarks for iterating on their models. For example, at the time of writing this article in September 2020, the Fivethirtyeight site gives a 95% predictive interval of (42%, 60%) for Biden's share of the two-party vote in Florida, and 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, at least to us. That is, the Florida interval seems too wide given that at the time of writing this article, Biden is 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 do not see how one could plausibly get to a Trump landslide in that state. In contrast, the California conditional prediction made by Fivethirtyeight seems too pessimistic on

Trump's chances: if the president really were to win that state, this would almost certainly happen in a Republican landslide (only Hawaii and Washington D.C. lean more toward the Democrats), 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 that we observe seem to reveal that the Fivethirtyeight forecast has a too-low correlation among state-level uncertainties. Their joint prediction 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.

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 achieve 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 popular 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. Other forecasters can and possibly do undertake such interrogations to fine-tune their models over time, both during election cycles and in between.

2.4 Visualizing uncertainty

There is a literature on communicating probability statements (for example, Gigerenzer & Hoffrage, 1995, Spiegelhalter, Pearson & Short, 2011) but it remains a challenge to express election forecasts so they will be informative to political junkies without being misinterpreted by laypeople. In communicating the rationale behind Fivethirtyeight's displays, 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 to deliver 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.

Given that probability itself can be difficult for laypeople to grasp, it will be especially challenging to communicate uncertainty in a complex multivariate forecast. One message from the psychology literature is that natural frequencies 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 so 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.

A growing body of work in computer science has proposed and studied static and dynamic visual encodings for uncertainty. While much of this work has focused on visualizing uncertainty in complex high dimensional data analyzed by scientists, some new uncertainty visualization approaches have been proposed to support understanding among broader audiences, several of which use a visual frequency framing. For example, animated hypothetical outcome plots (Hullman, Resnick & Adar, 2015) present random draws from a distribution over time, while quantile dotplots discretize a density into a set of 20, 50, or 100 dots (Kay et

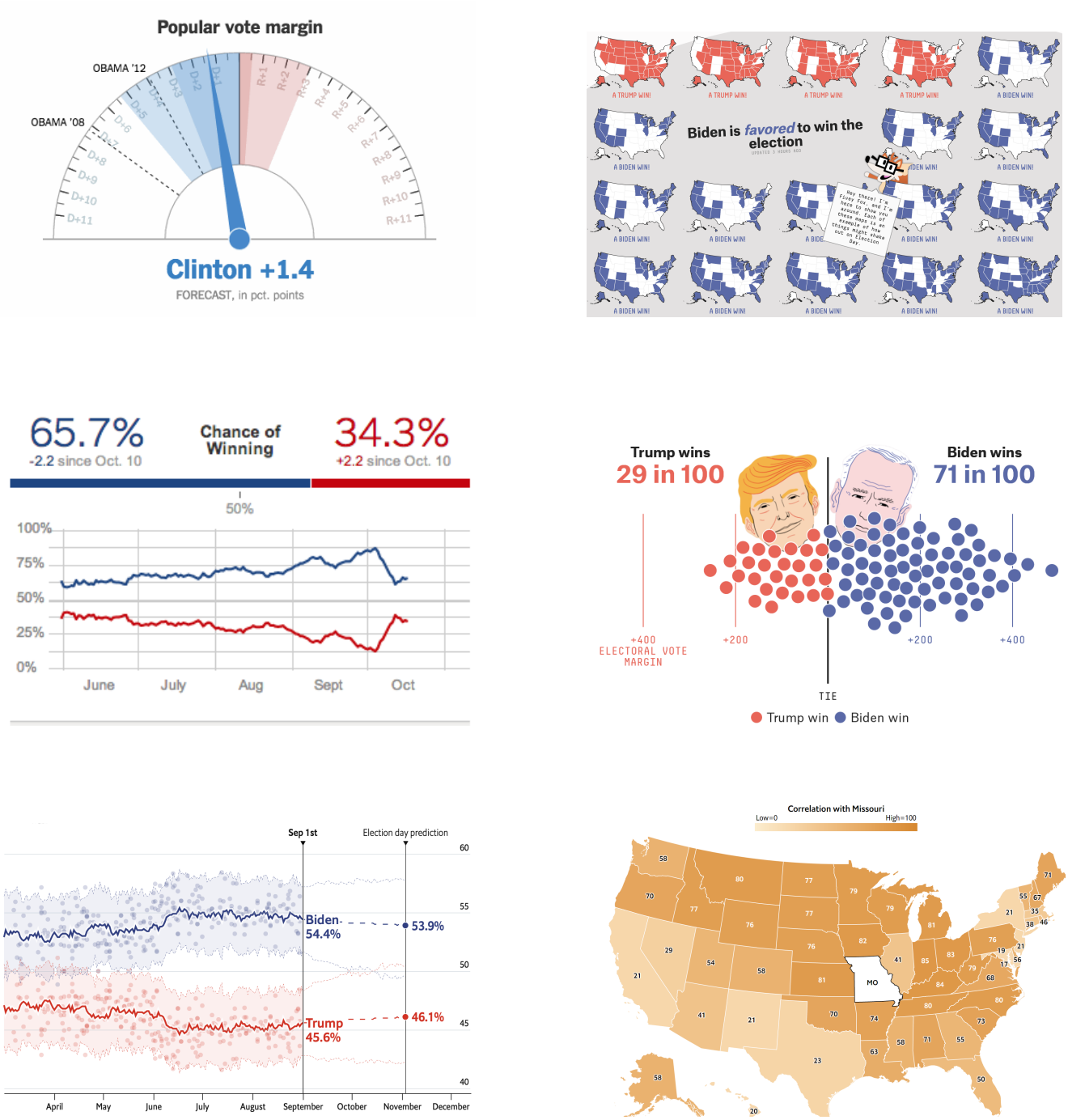

FIGURE 1: Some displays of uncertainty in presidential election forecasts. Top row: 2016 election needle from the New York Times and map icon array from Fivethirtyeight in 2020. Center row: time series of probabilities from Fivethirtyeight in 2012 and their dot distribution in 2020. Bottom row: time series of popular vote projections and interactive display for examining between-state correlations from the Economist in 2020. No single visualization captures all aspects of uncertainty, but a set of thoughtful graphics can help readers grasp uncertainty and learn about model assumptions over time.

al., 2016). Controlled experiments aimed at understanding how frequency-framed visualizations affect judgments and decisions made by laypeople provide a relevant base of knowledge for forecasters. For example, multiple studies compare frequency visualizations to more conventional displays of uncertainty. Several of these studies find these tools enable laypeople to make more accurate probability estimates and even better decisions as defined against a rational utility-optimal standard (Fernandes et al., 2018, Kale, Kay & Hullman, 2020) as compared to those who are given error bars and variations on static, continuous density plots.

Other studies suggest that the type of uncertainty information reported (for example, a predictive interval rather than a confidence interval) is more consequential in determining perceptions and decisions (Hofman, Goldstein & Hullman, 2020). One reason that it is challenging to try to generalize these findings from laboratory experiments is that people are likely to adapt various types of heuristics when confronted with uncertainty visualizations, and these heuristics can vary based on context. For example, when faced with a visualization showing a difference between two estimates with uncertainty, many people tend to look at the visual distance between mean estimates and use it to judge the reliability of the difference (Hullman, Resnick & Adar, 2015, Kale, Kay & Hullman, 2020). When someone is already under a high cognitive load, these heuristics, which generally act to suppress uncertainty, may be even more prevalent (Zhou et al., 2017).

There's even some evidence that people apply heuristics like judging visual distances to estimate effect size even when given what statisticians and designers might view as an optimal uncertainty visualization for their task. A recent study found that less than 20% of people who were given animated hypothetical outcome plots, which directly express probability of superiority (the probability that one group will have a higher value than another), figured out how to properly interpret them (Kale, Kay & Hullman, 2020). So even when a visualization makes uncertainty more concrete, it will also matter how forecasters explain it to readers, how much attention readers have to spend on it, and what they are seeking from the information.

Some recent research has tried to evaluate how the information highlighted in a forecast display — win probability or vote share — may affect voting decisions in a presidential elections (Westwood, Messing & Lelkes, 2020, Urminsky & Shen, 2019). One reason why studying the impact of visualization choices on voting decisions is challenging is that voting is an act of civic engagement more than an individual choice. Decision making in an economic or medical setting is more subject to probabilistic analysis because the potential losses and benefits there are more clear.

Figure 1 shows uncertainty visualizations from recent election campaigns that range from frequency based to more standard interval representations. The New York Times needle was an effective example of animation, using a shaded gauge in the background to show plausible outcomes according to the model, with the needle itself jumping to a new location within the central 50% interval every fraction of a second. The needle conveyed uncertainty in a way that was visceral and hard to ignore, but readers and journalists alike expressed disapproval and even anger at its use (McCormick, 2016). While it was not clear to many readers what exactly drove each movement of the needle, we think expectations were likely a bigger contributor to the disapproval: 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.

A more subtle frequency visualization is the grid of maps used as the header for Fivethirtyeight's forecasting page, with the number of blue and red maps representing possible combinations of states leading to a Biden or Trump win according to the probability assigned by the forecast. Visualization researchers have called grids of possible worlds "pangloss plots" (Correll & Gleicher, 2014), representing a slightly more complex example of icon arrays, which have long been used to communicate probabilities to laypeople in medical risk communication (Ancker et al., 2006). The Economist display designers also experimented with an icon-style visualization for communicating risk or "risk theater", which shaded a percentage of squares in a grid blue or red to reflect the percentage change that either candidate wins the electoral college.

For illustrating the time series of predictions during the campaign, the Fivethirtyeight lineplot is clear and simple, but, as noted in Section 2.2, 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 (Urminsky & Shen, 2019, Westwood, Messing & Lelkes, 2020).

Fivethirtyeight's dot distribution shows another frequency visualization. In contrast to the map icon array, the dot display also conveys information about how close the model predicts the electoral college outcome to be. Readers may be confused about how this particular set of 100 dots was chosen, and the display loses precision compared to a continuous display, but it has the advantage of making probability more concrete through frequency. Indeed, it was through these visualizations that we noticed the problematic state-level forecasts discussed in Section 2.3.

The Economist time series plot of estimated vote preference has the appealing feature of being able to include the poll data and the model predictions on the same scale. Here readers may be likely to judge how close the race is based on how far apart the two candidates' forecasts are from one another within the total vertical space of the y-axis (Kale,

Kay & Hullman, 2020), rather than trying to understand how close the two percentages are in other ways, such as by comparing to prior elections. Shaded 95% confidence intervals, which barely overlap in this particular case, help convey how sure the model is that Biden will win. If the intervals were to overlap more, people's heuristics for interpreting how "significant" the difference between the candidates is might be more error prone (Belia et al., 2005). The display does not map directly onto win probability or even electoral college outcomes and so may be consulted less by readers wanting answers, but, as discussed in Section 2.2, we believe that vote proportions are ultimately the best way to understand forecast uncertainty, given that short-term swings in opinions and votes tend to be approximately uniform at the national level. Electoral college predictions are made available in a separate plot.

Finally, the Economist's display includes an interactive choropleth map that allows the reader to select a state and view how correlated the model expects its voting outcomes to be with other states via shading. This map alerts readers to an aspect of election forecasting models that often goes unnoticed — the importance of between-state correlations in prediction — and lets them test their intuitions against the model's assumptions.

As in presenting model predictions in general, it is good to present multiple visualizations to capture different aspects of data and the predictive distribution as they change over time. Plots showing different components of a forecast can implicitly convey information about the model and its assumptions, and both Fivethirtyeight and the Economist do well by displaying many different looks at the forecast, along with supplying probabilistic simulations available for download.

2.5 Other ways to communicate uncertainty

It is 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 & Rappaport, 1986, Budescu, Weinberg & Wallsten, 1988).

For uncertainty that can't be quantified because it involves unknowns such as the credibility of assumptions, it may help to resort to qualitative text expressions like "there is some uncertainty around these results due to X." Some research suggests that readers take these qualitative statements more seriously than they do quantitative cues (van der Bles, Freeman & 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, as well as the implicit assumption that factors not included in the model are irrelevant to prediction. No model can include all such factors, thus all forecasts are conditional. We try our best to capture forecast uncertainty by calibrating the residual error terms on past elections, but every election introduces something new.

2.6 Prediction markets

A completely different way to evaluate forecasts and think about their uncertainties is to compare them to election betting markets. In practice, we would not expect such markets to have the martingale property; 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." Political betting markets, in particular, will respond to a series of new polls and news items throughout the campaign. The markets can overreact to polls or can fail in the other direction by self-reinforcing, thus paradoxically not making the best use of new data (Erikson & Wlezien, 2008, Gelman & Rothschild, 2016a). That said, markets offer a different sort of data than polls and fundamentals, and we should at least be aware of how these signals can disagree.

During the 2020 campaign, prediction markets in the 2020 campaign have consistently given Biden an implicit win probability in the 50–60% range, compared to poll-based forecasting models that have placed the Democrat's chance of winning to be in the 70–90% range. That said, this direct interpretation of the probabilities for winner-take-all prices is not entirely straightforward (Manski 2006).

This discrepancy between statistical forecasts and markets can be interpreted in various ways. The market odds can represent a mistake of bettors who are overreacting to the surprise outcome from 2016. Another possibility is that poll aggregation fails to account for systematic biases or the possibility of large last-minute opinion changes, in which case the markets could motivate changes in the forecasting models.

It is unclear exactly how we would incorporate betting market information into a probabilistic forecast. As noted, the markets are thin and, when state-level markets are also included, the resulting multivariate probabilities can be incoherent. We can think of the betting odds as a source of external information, alongside other measures such as reports of voter enthusiasm, endorsements, money raised, and features such as the coronavirus epidemic that could affect voter turnout in ways that are unique to the current election year and so are difficult to directly incorporate into a forecasting model.

Another reason that polls can differ from betting odds is that surveys measure vote intention, whereas the market applies to official vote totals; as discussed in Section 1.4, vote suppression and discrepancies in vote counts are not addressed by existing prediction models that use polls and fundamentals. In theory, and perhaps in practice, markets can include information or speculation about such factors that are not included in the forecasts.

3 The role of incentives

Election forecasting might 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 appear to be devoting more effort to better expressing uncertainty over time, as illustrated by the quote leading off Section 2.4 from Wiederkehr (2020), discussing choices made in displaying predictions for 2020 in response to criticisms of the ways in which forecasts had been presented in the previous election.

The acknowledgment that it can be risky to present numbers or graphs that imply too much precision may be a sign 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 issued by the news organization, 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. We discuss how these factors combine to make forecasters' incentives complex.

3.1 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 & Azari, 2017):

> There's a theory that academics such as ourselves are petrified of making a mistake, hence we are overcautious in our predictions; by 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 any such direct motivation for overconfidence, it is difficult for people to fully express their uncertainty when making probabilisitic predictions (Alpert & Raiffa, 1982, Erev, Wallsten & Budescu, 1994). 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.

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); feeling that there are no 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); and thinking that not presenting uncertainty will make it easier for people to coordinate beliefs (Manski, 2019).

There are also strategic motivations for forecasters to minimize uncertainty. 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 (for example Stevenson, 2016; Raza & Knight, 2020) with his "surefire guide to predicting the next president".

3.2 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 to report only 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, one of us 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 consumers embraced poll aggregation, forecasting became 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%s really mean 20%)" (Silver, 2020b). But making predictions conservative corresponds to increasing the widths of intervals, playing it safe by including extra uncertainty. Characterizing a forecasting procedure as conservative implies an attitude of risk-aversion, being careful to avoid the outcome of the predictive interval not including the actual election result. 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 their 20% really means 14%, and their 80% really means 88%. As discussed in Section 2.1, these numbers are based on a small number of elections so we shouldn't make too much of them, but this track record is consistent with Silver's goal of conservatism, leading 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. And Fivethirtyeight's probabilistic forecasts for sporting events do appear to be calibrated (Boice & 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 embarrassment 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 & LeClerc, 2012).

3.3 Incentives in competing forecasts

Incentives could get complicated if forecasters expect "dueling certitudes" (Manski, 2011), cases where multiple forecasters are predicting a common outcome. For example, suppose a forecaster knows that other forecasters will likely be presenting estimates that will differ from each other, at least to some extent. 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 fore-

caster 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 an altruistic 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.

Statisticians have examined how an expectation that forecasts will be combined and weighted often gives forecasters an incentive to deviate from reporting their true best prediction (Bayarri & DeGroot, 1989).. Economic literature on goals and strategies in forecasting can also shed some light on incentives in competitive environments. Here, it is assumed that forecasters have access to both public and private information sources, and that forecasters' behavior can be described through a mixture of concerns related to preserving their reputation for accuracy and maximizing their payoffs (Marinovic, Ottavani & Sorensen, 2013).

For example, the desire to avoid releasing information that could later be considered inaccurate leads forecasters to produce predictions closer to the consensus forecast than is warranted by the forecaster's private information. This is because the market is incentivized to separate the forecaster's private signal from the public prior to judge the quality of their information, but the forecaster is incentivized to incorporate the prior in the forecast. At equilibrium in such a game, the forecasters can truthfully communicate the direction of their private signals but not the intensities.

On the other hand, contest effects caused by a convex incentive scheme, where payoffs drop off significantly after the best forecast, lead to forecasts that overweight private information. This is because the forecaster wants to maximize the ratio of the probability of winning the contest (having the most accurate forecast), to the density of winning forecasts. The first order reduction in the expected number of winners (which is centered around the common prior) if the forecaster deviates from their true forecast toward their private signal is greater than the second order reduction in the probability of winning. Because the payoff is directly linked to the forecast, unlike in a reputation game, the incentive to deviate holds even in equilibrium (Ottavani & Sorensen, 2006, Marinovic, Ottavani & Sorensen, 2012).

Given how difficult it is to assess how well calibrated a forecast is in election forecasting, perceived reputational payoffs in the form of more reader attention or post-election praise for accuracy are likely more random, and may be subject to biases not considered by existing economic models. For example, if some market agents use heuristics such as a forecaster's personal characteristics or perceived political orientation to assess the quality of the forecaster's private signals, reputational payoffs may strengthen conservatism among some forecasters or even cause them to leave the market. The media environment tolerates failure from some, not all.

In the face of so much uncertainty about forecasters' performance in any single election, tendencies to exaggerate private signals may be even stronger than in the studied contest scenarios, as in the quotation in Section 3.1 in which Scott Adams claimed victory after predicting Trump would win in a landslide: in this case, getting the sign right overwhelmed any concerns about the estimated magnitude of the result. An attempt to apply formal models to election forecasting would undoubtedly lead to much weaker predictions about forecasters' optimal strategies than those that are possible for financial, sports, weather, or other domains with more frequent outcomes.

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% ± 2%; theirs is roughly 53% ± 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 1.3. There is no easy way to choose between 54% ± 2% and 53% ± 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 as much as its product.

3.4 Novelty and stability

There has been some discussion in the economic literature about how news organizations may display biases that systematically prioritize one party over another when they present political information like forecasts; for a review of supply and demand-side forces leading to biased political reporting in equilibrium see Gentzkow, Shapiro and Stone (2014). A less obvious challenge when producing forecasts 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 & Rothschild, 2016b, Kennedy et al., 2018). As discussed, events during the campaign can sometimes shift the fundamentals, but such events are rare (Erikson & 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. That said, as discussed in Section 2.2, 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 a continuing supply of "news", a forecast provides a baseline of expectation that allows us to interpret the real political news as it happens.

Again, all of the foregoing refers to the general election for president. Primary elections and other races, including for the U.S. House and Senate, can be much harder to predict and much more volatile, making forecasting a more challenging task with a greater 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 and organizations 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 has several implications:

- *Fundamentals*. Forecasters should be mindful of known regularities in election results and make use of information that research indicates has predictive power.
- *Data quality*. Polls have sampling and nonsampling error, and surveys that do not sufficiently adjust for differences between sample and population can have systematic biases.
- *State and national predictions*. Swings tend to be approximately uniform across the country, which implies that there is value in tracking national polls even for the goal of making a state-by-state electoral college forecast.
- *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.

At a high level, our work suggests that there are as many unknowns, in the form of evaluation challenges, in presidential election forecasting as there are knowns about how to create a proper forecast. By drawing attention to the difficulty of assessing calibration and the way this opens up space for forecasters' incentives to play a role, we hope to expand the typical public and scholarly discussion of forecast details to acknowledge a broader scope of issues.

As discussed in Section 1.6, none of the forecasts under discussion are fully Bayesian — at least in the generative sense — meaning that martingale properties of a Bayesian belief stream can't be expected to hold. Still, future attempts to formally validate election forecasting models might analyze them in terms of movement (how much a prediction varies over time) and uncertainty reduction given the net effects of information. More generally, the literature on expert testing such as Foster and Vohra (1998) or more recent work aimed at identifying optimal tests for forecaster quality (Deb, Pai & Said, 2018) may be useful for theorizing about calibra-

tion and incentives even in the absence of strong calibration data.

Responsibilities toward uncertainty communication are harder to outline. As discussed in Section 2.2, summaries such as win probabilities depend strongly on difficult-to-test assumptions, hence it is important for forecasters to air these 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 informed 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, forecasters have some responsibility to take into account what readers may do with a visualization or statement of forecast predictions. That readers 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 that 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: readers vary in their knowledge and interest in the topic, heuristics can look like accurate responses, and normative interpretations often don't exist (Hullman et al., 2018). However, this is not to say that principled communication of forecast uncertainty is not desirable. 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). While studying how different displays may affect voting behavior directly is challenging in the lab, researchers could help by pursuing empirically-trained models to inform decisions such as whether to acquire more information about a candidate's campaign, or make a campaign donation. For example, recent research on uncertainty visualization models decisions under uncertainty by estimating how people's "point of subjective equality", at which they are indifferent between two options or stimuli, shifts with different uncertainty displays (Kale, Kay & Hullman, 2020). Designing complex cognitive models to predict decision making from election forecasts may not be realistic given the heterogeneity of forecast consumers and available resources at a news organization, but designing a forecast without any thought to how it may play into readers' decisions seems both impractical and potentially unethical. In general, we think that more collaboration between researchers invested in empirical questions around uncertainty communication and journalists developing forecast models and their displays would be valuable.

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.

Naturally, adding too much information risks overwhelming readers. The majority spend only a few minutes on the websites, and may feel overwhelmed by concepts such as correlation that forecasters will view as both simple and important, but are largely beside the point of the overall narrative of the forecast. Still, increasing readers' literacy about model assumptions could happen in baby steps: a reference to a model assumption in an explanatory annotation on a high level graph, or a few bullets at the top of a forecast display describing information sources to whet a reader's appetite.

It may also 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 forecast uncertainty levels, particularly in settings

where assessing calibration is so difficult. We are aware of some academic discussions from economists and psychologists of incentives in constructing probabilistic forecasts (Marinovic, Ottaviani & Sorenson, 2012, Manski, 2011, 2018, Fischoff, 2012, Baron et al., 2014). In many cases, however, existing formulations make assumptions that do not necessarily hold in election forecasting. Still, we think that more such discussions are well motivated, if only to speculate about different possible scenarios for presidential election forecasters' incentives.

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 its 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 could not 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. (2010). *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy*. Yale University Press.

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.

Aldous, D. J. (2013). Using prediction market data to illustrate undergraduate probability. *American Mathematical Monthly, 120*, 583–593.

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In *Judgment Under Uncertainty: Heuristics and Biases*, ed. D. Kahneman, P. Slovic, & A. Tversky, 294–305. Cambridge University Press.

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association, 13*, 608–618.

Augenblick, N., & Rabin, M. (2018). Belief movement, uncertainty reduction, and rational updating. Haas School of Business, University of California, Berkeley. faculty.haas.berkeley.edu/ned/AugenblickRabin_MovementUncertainty.pdf.

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

Bayarri, M., & DeGroot, M. (1989). Optimal reporting of predictions. *Journal of the American Statistical Association, 84*, 214–222.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*, 389–396.

Boice, J., & Wezerek, J. (2019). How good are Fivethirtyeight forecasts? projects.fivethirtyeight.com/checking-our-work.

Budescu, D., Weinberg, S., & 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.

Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics, 20*, 2142–2151.

Cuzan, A. (2015). Five laws of politics. *PS: Political Science and Politics, 48*, 415–419.

Deb, R., Pai, M., & Said, M. (2018). Evaluating strategic forecasters. *American Economic Review, 108*, 3057–3103.

Enns, P., & Lagodny, J. (2020). Forecasting the 2020 electoral college winner. *PS: Political Science and Politics*.

Erev, I., Wallsten, T. S., & 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., & Stimson, J. A. (2002). *The Macro Polity*. Cambridge University Press.

Erikson, R. S., & Wlezien, C. (1999). Presidential polls as a time series: The case of 1996. *Public Opinion Quarterly, 63*, 163–177.

Erikson, R. S., & Wlezien, C. (2008). Are political markets really superior to polls as election predictors? *Public Opinion Quarterly, 72*, 190–215.

Erikson, R. S., & 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., & 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.

Fiorina, M. (2017). *Unstable Majorities: Polarization, Party Sorting, and Political Stalemate*. Hoover Institution Press.

Fischoff, B. (2012). Communicating uncertainty: Fulfilling the duty to inform. *Issues in Science and Technology, 28*, 63–70.

Foster, D., & Vohra, R. (1998). Asymptotic calibration. *Biometrika, 85*, 379–390.

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.

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-predict.

Gelman, A., & Azari, J. (2017). 19 things we learned from the 2016 election (with discussion). *Statistics and Public Policy, 4*, 1–10.

Gelman, A., Goel, S., Rivers, D., & Rothschild, D. (2016). The mythical swing voter. *Quarterly Journal of Political Science, 11*, 103–130.

Gelman, A., & 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., & Rothschild, D. (2016a). Something's odd about the political betting markets. *Slate*, 12 July. slate.com/news-and-politics/2016/07/why-political-betting-markets-are-failing.html.

Gelman, A., & Rothschild, D. (2016b). 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.

Gentzkow, M., Shapiro, J. M., & Stone, D. F. (2014). Media bias in the marketplace: Theory. National Bureau of Economic Research working paper 19880. www.nber.org/papers/w19880.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704.

Graefe, A., Armstrong, J. S., Jones, R., & 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., & 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., & 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, & 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. Sabato's Crystal Ball, 6 Aug. www.centerforpolitics.org/crystalball/articles/author/natalie-jackson.

Jennings, W., & Wlezien, C. (2016). The timeline of elections: A comparative perspective. *American Journal of Political Science, 60*, 219–233.

Joslyn, S., & 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., & Hullman, J. (2020). Visual reasoning strategies for effect size judgments and decisions. *IEEE*

Transactions on Visualization and Computer Graphics.

Kaplan, N., Park, D. K., & 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., & 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, & 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., & 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.

Lewis-Beck, M., & 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., & Gelman, A. (2010). Bayesian combination of state polls and election forecasts. *Political Analysis, 18*, 337–348.

Lohr S. L., & 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.

Manski, C. F. (2006). Interpreting the predictions of prediction markets. *Economics Letters, 91*, 425–429.

Manksi, C. F. (2011). Policy analysis with incredible certitude. *Economic Journal, 121*, F261–289.

Manksi, C. F. (2019). The lure of incredible certitude. *Economics & Philosophy, 36*, 216–245.

Marinovic, I., Ottaviani, M., & Sorensen, P. (2013). Forecasters' objectives and strategies. In *Handbook of Economic Forecasting*, volume 2B, ed. G. Elliott and A. Timmermann, 690–720. Elsevier.

McCormick, R. (2016). The NYT's election forecast needle is stressing people out with fake jitter. *The Verge*, 8 Nov. www.theverge.com/2016/11/8/13571216/new-york-times-election-forecast-jitter-needle.

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). Meet our US 2020 election-forecasting model. *Economist*, 11 Jun.

Morris, G. E. (2020b). How the Economist presidential forecast works. *Economist*, 5 Aug.

Morris, G. E. (2020c). More mail-in voting doubles the chances of recounts in close states. *Economist*, 22 Aug.

Ottaviani, M., & Sorensen, P. (2006). The strategy of professional forecasting. *Journal of Financial Economics, 81*, 441–466.

Page, B., & 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., & 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/2020-election-prediction-allan-lichtman.html.

Rosenstone, S. J. (1983). *Forecasting Presidential Elections*. Yale University Press.

Shirani-Mehr, H., Rothschild, D., Goel, S., & 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.

Silver, N. (2020b). Twitter thread, 1 Sep. twitter.com/NateSilver538/status/1300825871759151117.

Snyder, J., with Herskowitz, M., & Perkins, S. (1975). *Jimmy the Greek, by Himself*. Chicago: Playboy Press.

Spiegelhalter, D., Pearson, M., & 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.

Taleb, N. N. (2017). Election predictions as martingales: An arbitrage approach. *Quantitative Finance, 18*, 1–5.

Urminsky, O., & Shen, L. (2019). High chances and close margins: How equivalent forecasts yield different beliefs. ssrn.com/abstract=3448172.

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., & 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., & 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.

Westwood, S. J., Messing, S., & Lelkes, Y. (2020). Projecting confidence: How the probabilistic horse race confuses and demobilizes the public. *Journal of Politics, 82*, 1530–1544.

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.

Wlezien, C., & Erikson, R. S. (1996). Temporal horizons and presidential election forecasts. *American Politics Research, 24*, 492–505.

Wlezien, C., & Erikson, R.S. (2004). The fundamentals, the polls, and the presidential vote. *PS: Political Science and Politics, 37*, 747–751.

Zhou, J., Arshad, S. Z., Luo, S., & Chen, F. (2017). Effects of uncertainty and cognitive load on user trust in predictive decision making. *IFIP Conference on Human-Computer Interaction*, 23–39.