The Polls and the US Presidential Election in 2020 ….and 2024

Arguably, the single greatest determinant of US public policy is the identity of the president. And if trusted, polls not only provide forecasts about presidential-election outcomes but can act to shape those outcomes. Looking ahead to the 2024 US presidential election and recognizing that polls before the 2020 presidential election were sharply criticized, we consider whether such harsh assessments are warranted. Initially, we explore whether such polls as processed by the sophisticated aggregator *FiveThirtyEight* successfully forecast actual 2020 state-by-state outcomes. We evaluate FiveThirtyEight's forecasts using customized statistical methods not used previously, methods that take account of likely correlations among election outcomes in similar states. We find that, taken together, the pollsters and *FiveThirtyEight* did an excellent job in predicting who would win in individual states, even those "tipping point" states where forecasting is more difficult. However, we also find that *FiveThirtyEight* underestimated Donald Trump's vote shares by state to a modest but statistically significant extent. We further consider how the polls performed when the more primitive aggregator Real Clear Politics combined their results, and then how well single statewide polls performed without aggregation. It emerges that both Real Clear Politics and the individual polls fared surprisingly well.

Key Words: *US Elections, Political Polling, Forecast Accuracy, US Presidency*

Arnold Barnett[1] and Arnaud Sarfati[2]

1: George Eastman Professor, MIT Sloan School of Management, E62-568, MIT, Cambridge, MA 02142, abarnett@mit.edu, 617 686-1485 (corresponding author)

2: Student, Masters in Business Analytics program, MIT Sloan School of Management, MIT, Cambridge, MA 02142

To appear, *Statistics and Public Policy* (American Statistical Association)

1. Introduction

In 2020 as in 2016, there was widespread frustration concerning the performance of polls about the US presidential election. *The New York Times* ran an article titled “2016 Dealt a Blow to Polling. Did 2020 Kill It?” (S. Bokat-Lindell (2020)) Writing in *The Wall Street Journa*l, former Secretary of State James Baker declared that “this time, we were promised that pollsters would get it right (Baker (2020)). They didn’t.” A *Washington Post* op-ed [3] was titled “The Polling Industry Can’t Sweep Its Failure Under the Rug (Olson (2020), while *Fast Company* described 2020 as “another embarrassing failure for election pollsters (Campbell (2020)), and a commentator in Yahoo! Finance thought that “the biggest election takeaway (Ferre (2020) is the absolutely massive failure of polling.” The commentary about 2020 polls was just as harsh as that for their 2016 predecessors, even though, unlike those in 2016, the polls in 2020 correctly identified the winner of the election.

As the quotes above imply, whether presidential polls are accurate is an issue of immense importance in the United States. Arguably, the single greatest determinant of US public policy is the identity of the president. And if trusted, polls not only provide forecasts about presidential-election outcomes but can greatly affect those outcomes, playing a large role in the choice of each party’s presidential nominee and in the behavior of voters.

There is a large literature about how polls go beyond describing voter preferences and act to shape those preferences. One well-known phenomenon is the “bandwagon effect,” under which the fact that a candidate is ahead in the polls leads to support from some voters out of a desire to be on the winning side. There is also evidence that polling results can affect voter turnout: if the election does not appear close, then some people see no need actually to vote. There is also strategic voting based on polls, whereby citizens vote for candidates other than those they actually favor (e.g., they choose a lesser-desired candidate in a presidential primary because polls say that candidate would be stronger in the general election). If the polls are suspect, however, then all these behaviors could diminish. That might not be altogether a bad thing, but neither need it be an unalloyed good. Reliable polls that depict a close election can stimulate voter turnout, and some strategic voting can yield an election winner who best reflects the policy views of the majority of voters.

Because confidence in the polls is presumably tied to their recent performance, perceptions about how they fared the 2020 presidential election can, for better or worse, have consequences for the election in 2024. For that reason, this paper investigates the success or failure of the presidential polls prior to the election in 2020 elections.

Most negative assessments of 2020 polling relate to alleged deficiencies in local polls conducted in individual states. Yet even if these assessments are accurate rather than overwrought, it could be misleading to focus on the frailties of particular polls. The polls might perhaps more reasonably be treated as raw materials used by sophisticated aggregators who, taking account of the limitations of such polls as well as broader patterns, synthesize the polling results to devise probabilistic forecasts about what will happen in elections. If those predictions perform well, then the polls that contributed heavily to the forecasts might collectively be construed as successful despite their individual imperfections.

Pursuant to that viewpoint, we first concentrate here on aggregated forecasts and, more specifically, those advanced for the 2020 presidential election by *FiveThirtyEight*, which is arguably the best known and most respected of the aggregators. *FiveThirtyEight* went further than just predicting the winner of the election; it advanced a series of probabilistic assessments about state-by-state win/loss outcomes and the vote split in each state among the candidates (Trump, Biden, and third-party nominees). We evaluate these predictions using customized statistical methods, which go beyond those that *FiveThirtyEight* itself uses or that have appeared in recent literature about the accuracy of polling results. In actuality, we are initially evaluating the combination of the polls and *FiveThirtyEight* rather than the "raw" polls in themselves. If the combination succeeds, it is a joint success.

The accuracy of 2024 presidential polls is already a live issue at the start of 2023. Polls have already emerged about a rematch in 2024 between Joe Biden and Donald Trump. Other polls ask whether Democrats want Biden as their standard bearer in the 2024 election, and about how Trump would fare in 2024 against several possible Democratic opponents. Further polls ask Republicans whom they prefer as the party's nominee in the next presidential election. A political analyst for *New York* magazine noted that such presidential polls have "real world consequences," because they "affect the decision-making of potential candidates, operatives and activists." (Kilgore, 2023).

More specifically, if such polls—as distilled by a respected aggregator like *FiveThirtyEight*—are viewed as trustworthy, they could affect the intensity of pressure on Joe Biden to retire. They could influence Republican voters in state primaries who wonder whether Donald Trump could plausibly win reelection. The potential candidacies of Democrats like Amy Klobuchar or Republicans like Ron DeSantis could rise or fall with their standings in voter surveys. The polls, in other words, could play a sizable role in determining who each party's candidate will be. And once the nominees are chosen, polls could greatly influence media coverage of the election campaign. Indeed, it is routinely lamented that polls turn the election into a "horse race," in which who is ahead and by how many furlongs gets greater attention that what the candidates say about the issues.

Moreover, there is reason to fear that, as in 2020, questions whether warranted or not will be raised about the legitimacy of the 2024 election outcome. Pre-election polls that are trusted can cast light on the credibility of such accusations. The inverse of these statements is also true: if the polls are not taken seriously, they cannot help adjudicate controversies about the 2024 election.

Here we restrict ourselves to the statistical accuracy of 2020 US presidential polls, and not to broader issues about the proper role for polling in the selection of the president. And, as suggested, we proceed on the premise that the polls' performance in the most recent presidential election is the best single indicator of their ability to answer the questions about 2024 that motivate them.

As we will see, the FiveThirtyEight-mediated forecasts about the 2020 elections fared well, the only shortcoming being a modest underprediction of Donald Trump's state-by-state vote shares. To gauge the centrality of FiveThirtyEight's own statistical modeling to that favorable outcome, we then turn to the corresponding forecasts from the less sophisticated but highly influential aggregator Real Clear Politics, which simply averages recent polls together with no attempt to correct for their potential biases. Then we step back from aggregated forecasts, to consider the heavily-criticized results from original local polls. To a surprising degree, we find that both Real Clear Politics and the original polls did well in their own right.

1.1 Previous Work

There is a large literature that suggests that pre-election polls affect election outcomes. Based on an experiment, Farjam (2020) discerned a substantial bandwagon effect, estimating that "after participants saw pre-election polls, majority options on average received an additional 7% of the votes." (Farjam also offers an extensive bibliography of papers about polls and elections.) Burden (2005) explored strategic voting in US presidential elections with respect to supporters of third-party candidates, concluding that many such supporters shifted their votes in 2000 to the "lesser evil" between the two major candidates, but that they rarely did so in 1992 and 1996 when polls suggested an easy victory for Bill Clinton. Bursztyn et al (2017) estimated that voter turnout increased when polls indicated a close race (with the implication that, when polls depicted a race that was not close, turnout declined relative to an average election). Westwood, Messing, and Lelkes (2020) lamented that "probabilistic horse race" election coverage—like that advanced by *FiveThirtyEight* based on pre-election polls--"confuses and demobilizes" the public, and concluded that confidence among her supporters in 2016 that Hillary Clinton would win the presidency was associated with lesser voter turnout.

There is also a voluminous literature about the accuracy of political polls. It is useful to distinguish those evaluations based on the individual polls and those that concern the forecasts of aggregators like *FiveThirtyEight*, which combine various polling results having adjusted for shortcomings among the polls. Below we turn first to some papers in the former category (polls only).

Arneson and Bergfjord (2014) studied US presidential polls in the US in 2008 and 2012, and offered evidence that their estimates about probabilities of victory were further from the mark than were chances of winning derived from the odds in betting prediction markets.

Prosser and Mellon (2018) questioned whether conspicuous failures to predict the winners in recent US and UK elections had created a "twilight of the polls." They discussed several reasons that polls fell short, including late swings, inadequate turnout models, mishandling of undecided voters, and unrepresentative samples sometimes tied to nonresponse biases. In the 2016 US presidential election, the authors cited the failure to weight properly for voter education levels as contributing to the underestimation of Donald Trump's strength. However, the authors concluded that polls were not getting worse over time, and thus it was excessive to suggest their imminent demise.

Panagopoulos (2021) discerned systematic polling errors in the 2020 US election cycle, which he said reflected "pro-Democratic biases." This pattern appeared in national and state-level polls in races for president, US Senate, and state governors. Panagopoulos saw growing difficulties tied to rising costs, declining response rates and "a host of technical and methodological challenges" that pollsters need to confront. "In the meantime," he advises, "the public is wise to consume polling information with caution."

Of exceptional importance in considering pre-election polls are the "post mortems" performed by the American Association for Public Opinion Research (AAPOR). The association's evaluation for the 2016 presidential race was less negative than one might expect, given the widespread shock at Trump's victory over Clinton (Ad Hoc Committee (2017)). The statewide polls, AAPOR concluded, correctly indicated a competitive, uncertain contest and only implied that Clinton was ahead by "the slimmest of margins." What weakened the polls according to AAPOR was an overrepresentation of college graduates, and the fact that undecided voters broke heavily for Trump. Moreover, forecasts about turnout were seemingly off, perhaps because some Clinton supporters, treating her election as a foregone conclusion, saw no need actually to vote.

AAPOR was somewhat more critical of presidential polls in 2020 (Clinton et al (2021)). It described polling errors as "of unusual magnitude" and saw a tendency to overstate Biden's vote shares in individual states and to understate Trump's, a tendency that was greater in the states that supported Trump in 2016. Yet AAPOR did not see a repetition of the problems that it had noted in 2016: college graduates were not overrepresented in the surveys, and late deciding voters split evenly between Trump and Biden. Furthermore, contrary to some theories, the Association found those Trump supporters who participated in polls were not reluctant to declare their preference. AAPOR felt unable to explain why the polls faltered in 2020, though it speculated that voters supporting Trump were less willing to speak with surveyors than those who opposed him. (In 2020 as in many previous years, only a minority of those contacted by pollsters agreed to take part in the canvassing.)

What AAPOR did not do was to consider the possibility that a polling aggregator like *FiveThirtyEight* was aware of the biases in presidential surveys in a given year and had largely corrected for them. Whether that happened in 2020 is a major focus of this paper.

As for *FiveThirtyEight* itself, several papers have addressed its performance in presidential elections prior to 2020. Barnett (2018) spoke favorably of *FiveThirtyEight's* record in the 2016 election, noting that it estimated that Trump had about a 30% chance of winning and praising its awareness that outcomes in Pennsylvania, Michigan, and Wisconsin—the three key states Clinton was expected to carry but which went to Trump—were positively correlated.

Two other performance reviews for *FiveThirtyEight* were less unabashedly positive. Wright and Wright (2018) explored *FiveThirtyEight's* state-by-state record in the 2016 presidential elections. The authors acknowledged that *FiveThirtyEight* had treated a Trump victory as only moderately unlikely, but suggested that the website had paid insufficient attention to a late-developing trend towards Trump. They advanced a smoothing mixed-effects model sensitive to both national and local trends that they argued would have performed better than *FiveThirtyEight* in 2016.

Rothschild (2009) evaluated *FiveThirtyEight* in connection with the 2008 presidential election, the first in which FiveThirtyEight offered forecasts. While he had favorable things to say about *FiveThirtyEight,* he concluded that the website suffered from some anti-incumbency bias, meaning a tendency to underestimate incumbents' vote shares. Rothschild suggested than a reason for this bias could be understating the extent to which voters who declare themselves undecided to pollsters vote for the incumbent on Election Day. He conducted a comparison between *FiveThirtyEight* and betting prediction markets and argued that, while *FiveThirtyEight* offered more accurate election forecasts, its advantage disappeared when the forecasts by prediction markets were "debiased."

However, *FiveThirtyEight* presumably has sought to improve its forecasting techniques over time based on any shortcomings it identifies. For that reason, its success or lack thereof in presidential elections before 2020 bears an unknown relationship to its performance in 2020 itself.

Nate Silver, the founder of *FiveThirtyEight,* himself writes after each presidential election about the accuracy of its forecasts. He concluded (Silver (2020)) that its predictions in the 2020 Biden/ Trump race "did very well." Silver drew attention to what he called the "rigorous methods" that *FiveThirtyEight* uses to evaluate its own performance, which we will discuss at length in Appendix B.

2. Materials and Methods

2.1 FiveThirtyEight

We focus on the website *FiveThirtyEight* created by Nate Silver because it is probably the best known and arguably the most respected among election-forecast aggregators in the United States. We concentrate on its *final* state-by-state predictions for the 2020 presidential election released in early November (FiveThirtyEight 2020) and which, unlike its earlier forecasts, give scant weight in key states to economic and historical factors and are based almost exclusively on polls conducted within the state[1] . The website takes a weighted average of polls, related both to their recency and to their patterns of error in recent forecasts. If a poll has tended systematically in the past to (say) overstate the actual vote shares of Republican candidates, *FiveThirtyEight* applies a correction for that bias. To some extent, the projections consider possible correlations among the outcomes in similar states. We consider the accuracy of predictions about Donald Trump's performance but, because Trump and Joe Biden were essentially in a two-person race, the analysis of Biden's performance would yield

[1] In FiveThirtyEight's nine swing states most likely to "tip" the election, the median weight it accorded to polls in its final 2020 forecast was 97%.

equivalent results. (For actual election results, we turn to the Federal Election Commission (2021).)

For a given state, *FiveThirtyEight* presents:

- An estimate of the probability that Trump will win that state
- A point estimate of Trump's share of the popular vote
- An 80% confidence interval for Trump's share of the popular vote, which extends from the 10th percentile to the 90th percentile of *FiveThirtyEight's* distribution for that quantity. The point estimate is at the midpoint of the confidence interval.

Actually, there are 56 "states" according to *FiveThirtyEight*: the usual 50 states, plus the District of Columbia, the three congressional districts in Nebraska, and the two in Maine. (In these two states, the popular-vote winner in a congressional district gains its Electoral-College vote regardless of the statewide outcome.)

We will conduct tests of the accuracy of FiveThirtyEight's 2020 projections by state. We believe our approach to assessing its predictive accuracy is more stringent and transparent than the validation procedures the website itself uses.

2.2 Win/Loss Projections

The simplest question one might ask about *FiveThirtyEight's 2020* performance is: how many states did it get right? By "right," we mean that the website assigned the winner a victory-probability higher than ½. An equivalent question is: how many states did the website get wrong? This right/wrong dichotomy lacks any nuance: if a candidate assigned a 45% chance of winning actually does so, then declaring the forecast an error seems superficial. But one can compare the website's actual number of "erroneous" forecasts with the number implied by its probabilistic projections.

Let P_{Li} be *FiveThirtyEight's* estimated probability that the *disfavored* candidate in state i actually wins (meaning that $P_{Li} < \frac{1}{2}$) and let random variable Z be the number of "erroneous" forecasts over the 56 states. Then the website's mean number of errors would follow:

$$E(Z) = \sum_{i=1}^{56} P_{Li}$$

If the outcomes in different states were assumed independent, then the variance of the number of errors would be given by:

$$\sigma^2(Z) = \sum_{i=1}^{56} P_{Li}(1 - P_{Li}) \qquad (1)$$

However, the outcomes across states may not be independent, a circumstance we will discuss.

2.3 The "Tipping Point" States

An assessment of *FiveThirtyEight'*s accuracy in 2020 should presumably give major emphasis to its performance in nine pivotal states, which the website identified as those "close to the tipping point" where the election would likely to decided. These nine states are:

Arizona Florida Georgia Michigan

Minnesota North Carolina Nevada Pennsylvania Wisconsin

Table 1 presents *FiveThirtyEight'*s estimate of Trump's chance of winning in each of the tipping point states.

Table 1: *FiveThirtyEight 's* Nine Tipping States in the 2020 Presidential Election,

State	Electoral Votes[a]	Estimated Probability of Trump Victory
Arizona	11	32%
Florida	29	31%
Georgia	16	42%
Michigan	16	5%
Minnesota	10	4%
North Carolina	15	36%
Nevada	6	12%
Pennsylvania	20	16%
Wisconsin	10	6%
Total	133	

Notes:

a: In US elections, the statewide winner gets all its electoral votes. There are 538 electoral votes across the United States; a candidate who gets at least 270 of them wins the election.

Let random variable S be the total number of swing states Trump would carry. We can approximate the probability mass function for S, based on both *FiveThirtyEight*'s "win" probability for Trump in each state and the estimated correlations of outcomes across the states.

We define the indicator variable X_i for each of the nine listed states by:

$$X_i = \begin{cases} 1 \; if \; Trump \; wins \; state \; i \\ 0 \; if \; Trump \; loses \; state \; i \end{cases}$$

The i's reflect alphabetical ordering of the states, meaning that X_1 refers to Arizona, etc.

Then the total number S of Trump wins would follow:

$$S = \sum_{i=1}^{9} X_i$$

Then, according to *FiveThirtyEight* just prior to the election, the mean of S would be given by:

$$E(S) = \sum_{i=1}^{9} E(X_i) = \sum_{i=1}^{9} p_{Ti} \qquad (2)$$

where $\mathrm{p_{Ti}} = \mathrm{P}(\text{Trump would win state i according to } \mathit{FiveThirtyEight})$

2.3.1 Correlated Outcomes Across Tipping Point States

In estimating the variance and standard deviation of S, we need consider that the election outcomes in different states can be correlated. For example, two states that had the same winner in all presidential elections from 1976 to 2016 would seem likely to go the same way again. We have the general expression:

$$\sigma^2(S) = \sum_{i=1}^{9} \sigma^2(X_i) + 2 \sum_{1 \le i < j \le 9} Cov(X_i\,, X_j\,). \qquad (3)$$

To estimate the covariance Cov(X_i , X_j), we focus on A_{ij} [2], the proportion of times the two states supported the same candidate in the presidential elections between 1976 and 2016. We initially set out four linear equations as follows

[2] FiveThirtyEight's modeling allows for correlated forecasts across states, but it does not disclose how, and our own approach to correlation is probably different. But test of a model need not be predicated on treating all its assumptions as correct (e.g., someone evaluating a model that assumes the earth is flat is not required to do likewise).

$$P_{TB} + P_{TT} = p_{Ti}$$

$$P_{BT} + P_{TT} = p_{Tj} \qquad (4)$$

$$P_{BB} + P_{TT} = A_{ij}$$

$$P_{TB} + P_{TT} + P_{TB} + P_{BB} = 1$$

where $P_{TT} = P(Trump\ carries\ both\ states)$

$$P_{TB} = P(Trump\ carries\ state\ i\ but\ not\ state\ j)$$

$$P_{BT} = P(Trump\ carries\ state\ j\ but\ not\ state\ i)$$

$$P_{BB} = P(Trump\ loses\ both\ states\ to\ Biden)$$

$$p_{Tk} = FiveThirtyEight's\ estimate\ of\ the\ chance\ that\ Trump\ will\ win\ state\ k$$

$A_{ij} = Fraction\ of\ presidential\ elections\ over\ 1976 - 2016$ *with the same outcome in states i and j*

The first two of these linear equations equate Trump's chance of winning a given state in 2020 to *FiveThirtyEight*'s probability estimate for that outcome. The third equation uses the frequency with which states i and j agreed on presidential outcomes over 1976-2016 as an estimate of the chance they would agree again in 2020.

But there is a potential problem. The four linear equations for four unknowns in (4) can be solved for $P_{TT}, P_{TB}, P_{BT}, and\ P_{BB,}$ but there is no guarantee that these quantities will all fall in the range (0,1). To avoid that problem, we pull back on the requirement that $P_{BB} + P_{TT}$ must equal A_{ij} and insist instead that $P_{BB} + P_{TT}$ be as close as possible to A_{ij} in a least-squares sense, consistent with feasible solutions for the various probabilities. We do so by advancing the following optimization model in quadratic mathematical programming:

$$\text{Minimize. } (P_{BB} + P_{TT} - A_{ij})^{2}$$

Subject to:

$$P_{TB} + P_{TT} = p_{Ti}$$

$$P_{BT} + P_{TT} = p_{Tj} \qquad (5)$$

$$P_{TB} + P_{TT} + P_{TB} + P_{BB} = 1$$

$$P_{TB},\ P_{TT}, P_{TB},\ P_{BB} \geq 0$$

Note that, whenever (4) yields a solution with non-negative probabilities, it is also the solution to (5).

Once we have the estimate of P_{TT} from (4), we can reach the probability distribution for Z_{ij}, which is defined by:

$$Z_{ij} = \begin{cases} 1\ if\ Trump\ carries\ states\ i\ and\ j \\ 0\ otherwise \end{cases}$$

Note that Z_{ij} is the product of $X_i\ and\ X_{j\ .}$, meaning that $E(Z_{ij}) = E(X_iX_j)$.

Note too that $E(Z_{ij}) = P_{TT}$.

We then have:

$$Cov(X_i, X_j) = E(Z_{ij}) - E(X_i)E(X_j) = P_{TT} - p_{Ti}p_{Tj} \quad (6)$$

Using the covariances calculated via (5) and (6) for the $\binom{9}{2} = 36$ combinations of i and j and noting that $\sigma^2(X_i) = p_{Ti}(1 - p_{Ti})$, we can obtain via (3) an estimate of $\sigma^2(S)$.

However, knowing the mean and standard deviation of S does not immediately yield its probability mass function, which takes the form:

$$S\ = j\ \ with\ probability\ q_j \quad for\ j = 0, 1, \dots 9$$

Consistent with the mean and standard deviation calculated for S, the $q_j's$ must satisfy three linear equations:

$$\sum_{j=0}^{9} jq_j = E(S)$$

$$\sum_{j=0}^{9} j^2\, q_j \;=\; E(S^2) = \;\sigma^2(S) + (E(S))^2 \qquad (7)$$

$$\sum_{j=0}^{9} q_j = 1$$

But there are ten $q_j's$, and these equations impose only three constraints on them. In consequence, there are many feasible sets of $q_j's$ that satisfy (7).

As described in Appendix A, we use an algorithm to obtain a "composite" distribution for S, in essence averaging across the feasible distributions consistent with (7).

Once the composite distribution for S is at hand, one can see where the actual number of tipping states that Trump won falls within that distribution. If it falls at (say) the extreme right tail, then the accuracy of *FiveThirtyEight*'s projections about the tipping states would be called into question.

2.4 *Trump's Vote Share*

UCLA Coach Henry Russell Sanders informed his players that "winning isn't everything; it's the only thing." But *FiveThirtyEight* accompanied its "win/loss" probabilities with probability distributions for the proportion of the vote Trump would receive in each state. For a full test of *FiveThirtyEight's* prediction methodology, it is important to explore how well those vote-share forecasts fared against the actual Trump/Biden vote split.

Specifically. *FiveThirtyEight* offered 80% confidence intervals for Trump's vote share in each of the 51 states, with the point estimate at the center of the interval. (We exclude the congressional districts of Maine and Nebraska from this analysis, because their vote-share data are fully contained in the statewide data we use.) Assuming normal distributions, the confidence interval ranged from the 10th percentile to the 90th percentile, namely, from 1.28 standard deviations below the mean to 1.28 standard deviations above it. Therefore, if the two bounds were a_{10} and a_{90} in a given state, the corresponding mean μ and standard deviation σ would be given by:

$$\mu = \frac{a_{10} + a_{90}}{2}$$

$$\sigma = \quad (a_{90} - \quad \mu)/1.28. \quad = \quad (a_{90} - a_{10})/2.56$$

We consider the null hypothesis H_0 that all of *FiveThirtyEight*'s probability distributions were accurate, meaning that the actual result in each state was one random pick from the state's specified normal distribution. Let random variable W_i be that pick when expressed as *a percentile* from the website's normal distribution for state i. Under H_0, W_i would be U(0, 100), because an outcome between (say) the 7th and 8th percentiles would have the same 1% chance of arising as one between the 43rd and 44th, or the 82nd and 83rd. For that reason, the mean of W_i would be 50 under H_0, and the standard deviation of W_i would be 29, based on general properties of uniform distributions.

A test of H_0 could fruitfully focus on $\overline{W}$, the arithmetic average of the 51 W_i's. The two-sided p-value associated with $\overline{W}$ would follow:

$$p - value = \begin{cases} 2 * P(R \ \geq \ \bar{W} \ ) \, if \ \overline{W} \ \ \geq 50 \\ 2 * P(R \ \leq \ \overline{W} \ ) \, if \ \overline{W} \ \ < 50 \end{cases}$$

where R is the average of 51 (correlated) picks from the joint distribution of the W_i's

One would reject H_0 if the p-value falls below some threshold value, the most common of which is .05.

If the different W_i's were independent random variables, then $\overline{W}$ under H_0 would be approximately normally distributed given the Central Limit Theorem, with a mean of 50 and a standard deviation of $\frac{29}{\sqrt{51}} = 4.07$. But, like the X_i's in the tipping states, the W_i's need not be independent, because high-percentile outcomes in some states could foreshadow similar outcomes in others. We will consider this point in connection with actual results.

3. Results

3.1 Win/Loss Forecasts

As discussed, *FiveThirtyEigh*t's number of state-by-state outcomes in which the disfavored candidate would win (X_L) has a mean μ_L which follows:

$$\mu_L = \sum_{i=1}^{56} p_{iL}$$

where p_{iL} = P(disfavored candidate wins in state i according to *FiveThirtyEight*). (p_{iL} < ½)

Based on *FiveThirtyEigh*t's 56 win/loss probabilities for 2020, $\mu_L = 5.17$. (All data we used about *FiveThirtyEight*'s final presidential forecasts for 2020 appear in FiveThirtyEight(2020).) That statistic shows that *FiveThirtyEigh*t was not timid in projecting winners: because 5.17 is less than 10% of 56, the website expected that it would correctly identify the winner more than 9/10 of the time.

In actuality, only three of *FiveThirtyEigh*t's win/loss forecasts were incorrect (in Florida, North Carolina, and the Northern Congressional District of Maine). Based on independence, the standard deviation of the number of errors (Z) would be 1.92 under (1). Furthermore, simulation reveals that the outcome three would be at the 24th percentile of the distribution of Z (i.e., the simulated value of Z was three or lower 24% of the time). Correlations between pairwise outcomes in different states could affect $\sigma^2(Z);$ however, their effect is unlikely to imperil the conclusion that the observed error rate fell comfortably in the distribution for the error rate based on FiveThirtyEight's state-by-state assessments[3] . In any event, when the estimated win/loss success rate is 91% (50.83/56) and the actual rate is 94% (53/56), there is no credible basis for claiming that *FiveThirtyEigh*t overestimated the accuracy of its win/loss predictions. The website made sharp claims about what would happen and satisfied them.

3.2 Trump's Success in the Tipping States

Table 1 offered some details about the nine states that *FiveThirtyEight* deemed most likely to "tip" the presidential election to the winning candidate. The issue is how closely the website's predictions about those states as a group corresponded to what actually happened in 2020.

In section 2.3.1, we described how we estimated the $\binom{9}{2} = 36$ correlation coefficients for the various pairings among the nine tipping states. Once those correlations and thus the

[3] Suppose that state-by-state win/loss outcomes for Trump are positively correlated. Then a Trump victory in state A could moderately increase his chance of winning in state B, where FiveThirtyEight assigns him a 25% chance of victory, and also do so in state C (75% chance). But then the conditional probability of a win/loss error would go up from 25% at B, but this error probability would go *down* from 25% at C. Thus, relative to independence, the net effect on $\sigma^2(Z)$ of these two opposite movements could well be modest.

$Cov\left(X_i, X_j\right)'s$ are at hand, the mean and variance of the number of tipping states S that Trump would win can be estimated based on (1), (2), and Table 1. The results are:

$$E(S) = 1.82 \qquad and \quad \sigma^2(S) \approx 1.64$$

We describe E(S) here as exactly 1.82 because it is based literally on the probabilities from *FiveThirtyEight* that appear in Table 1.

The next issue is the distribution of discrete random variable S, the number of the nine tipping states that Trump would win. S is a discrete random variable with a probability mass function of the form:

$$S \ = j \ \ with\, probability\ q_j \quad for\ j = 0, 1, \ldots 9$$

We describe our procedure for making estimates of the q_j's in Appendix A. We reach the following distribution:

$$S \ = \begin{cases} 0\ w.p..094 \\ 1\ w.p..328 \\ 2\ w.p..380 \\ 3\ w.p..118 \\ 4\ w.p..041 \\ 5\ w.p..018 \\ 6\ w.p..010 \\ 7\ w.p..005 \\ 8\ w.p..004 \\ 9\ w.p..002 \end{cases} \qquad (8)$$

w.p. = with probability

Among the nine states of Table 1, Trump actually won in two of them (Florida and North Carolina). That outcome is at the mode of S in (8) and as close as possible to *FiveThirtyEight*'s hypothesized mean of 1.82. In short, it appears that *FiveThirtyEight* did an excellent job in the tipping states, and that to treat Trump's victories in Florida and North Carolina as "errors" by the website requires valuing binary “win/lose” variables above the more nuanced use of actual probabilities and correlations.

Under *FiveThirtyEight* probability assessments*,* Trump’s bleak prognosis in the tipping states all but guaranteed his defeat. The remaining 47 states (56 – 9) included 22 for which the website favored Biden and 25 for which it favored Trump. But based on state-by-state win probabilities in the 47 states, Biden would gain a mean of 245 electoral votes, while Trump’s mean gain would be 160 electoral votes. To counter that mean difference of 85 (245-160), Trump would needed at least 109 of the 133 electoral votes in the tipping states[4] . Table 1 implies that doing so would have required at least seven Trump victories in the nine states, an outcome that is assigned a probability 0.011 in (8). Even that probability is an upper bound, because many combinations of seven or eight Trump victories would fall short of yielding 109 electoral votes (e.g., all those that exclude Florida).

3.3 Vote-Share Distributions

As we discussed in Section IV, our test of the accuracy of *FiveThirtyEight’s* state-by-state Trump vote share distributions entailed expressing his 2020 vote shares in individual states as percentiles of the corresponding *FiveThirtyEight* distributions. The observed outcomes in the 2020 election tilted decisively toward the upper tails of those distributions, with the average percentile over the 51 states at 69.52% . Under H_0 (all *FiveThirtyEight* distributions are correct), that outcome is 4.75 standard deviations above the expected value of 50% if the

[4] While the projected Biden/Trump difference in electoral votes could fluctuate around its mean of 85 for these 47 states, that circumstance would not meaningfully alter this approximate analysis.

standard deviation is estimated as 4.07% (i.e., assuming independence across states), meaning that the calculated p-value would be infinitesimal.

However, the presence of (generally positive) cross-state correlations would increase the standard deviation of $\overline{W.}$ Data that would directly allow estimating the correlation between Trump's vote share in state i (as a *FiveThirtyEight* percentile) and his share in state j are not available. However, generalizing the method we applied for the nine tipping states, we can estimate the correlations of Trump's binary win/loss variables for all $\binom{51}{2} = 1275$ pairs of states, and thus *FiveThirtyEight's* overall standard deviation for T, the total number of states Trump would carry (which was assigned a mean of 23.2). That standard deviation was 3.66, about twice the standard deviation of T assuming independence, which was 1.79. Using that two-to-one ratio as an approximate guide to what would happen to $\sigma(\overline{W\ })$ because of correlation, we might double its value as calculated based on both independence and H_0. Then the observed value of 69.52% would still be about 2.4 standard deviations above the mean, with a two-sided p-value of 0.0164. Thus, the null hypothesis H_0 that *FiveThirtyEight*'s 51 Trump vote-share distributions were all accurate would again be rejected at the usual 5% significance level.

That this adverse outcome is not spurious is further supported by the fact that in 47 of the 51 states—all except Alaska, Colorado, Louisiana, and Maryland--*FiveThirtyEight* underestimated Trump's support on election day. Over the 51 states, *FiveThirtyEight*'s point estimates of Trump's vote share were too low by an average of 1.90 percentage points. (Both *FiveThirtyEight's* projections and actual vote tallies took account of third parties, which received about 2% of votes nationwide.). Trump outperformed the point estimates in both heavily

Democratic states like Delaware, Massachusetts, and New York and heavily Republican states like North Dakota, Kentucky, and Wyoming. Interestingly, Trump exceeded his projected vote share by twice as much in states that he won as in states that he lost (2.56 percentage points versus 1.26). And when the state-by-state difference between Trump's vote share and FiveThirty Eight's point estimate for that share is regressed via OLS on the explanatory variable "point estimate," a statistically significant positive slope emerges:

$$y = 0.055x - 0.707$$

$y = difference\ between\ Trump's\ actual\ 2020\ vote\ share\ and\ FiveThirtyEight\ point\ estimate (\%)$
$x = FiveThirtyEight\ point\ estimate\ of\ Trump's\ vote\ share\ (\%)$

$Slope\ standard\ error: 0.018.$ $\qquad p - value\ of\ slope\ estimate:\ 0.004$
$Correlation\ of\ x\ and\ y: 0.401$

In essence, the more favorable to Trump's performance was *FiveThirtyEight*'s estimate, the greater in general was the extent to which it was not favorable enough.

Given that *FiveThirtyEight* underestimated Trump's vote share by a statistically-significant average of about two percentage points, why was it so successful in forecasting the winners in individual states? The answer is that, in the vast majority of states, the vote shares of Trump and Biden differed by more than two percentage points. That Trump carried Idaho with 63.8% of the vote rather than the projected 59.5% had no effect on the state's win/loss outcome; that Biden carried Hawaii with 65.7% of the vote rather than 69.1% was likewise immaterial.

3.4. Polling Accuracy Without FiveThirtyEight's Intervention

We have discussed the wide public perception that the presidential polls failed in 2020, and the detailed negative conclusions reached by experts at AAPOR. For that reason, we have been treating the individual 2020 polls as presumptively deficient and exploring whether

FiveThirtyEight counteracted their weaknesses. But it is worth estimating how large was the accuracy problem that *FiveThirtyEight* was meant to ameliorate.

It is helpful in this connection to turn to Real Clear Politics, which offers direct information about the collective performance of pre-election presidential polls. In each state where Real Clear Politics (hereafter RCP) saw the 2020 election as close, it simply took an arithmetical average of local polling results for a limited period before election day. It paid no attention to the sample sizes of individual polls, to a given poll's recency (e.g., two weeks before the election or three days before), or to any poll's historical tendency to favor one political party over the other. If one worked with the RCP averages alone, how much worse would have been the forecasts than those produced by *FiveThirtyEight*?

To answer that question, one first has to make reasonable adjustments to the raw RCP data. For example, suppose that RCP's last estimate before the election that Trump and Biden were tied at 46% in a given state, with 5% undecided and 3% favoring third-party candidates. Then RCP was not actually predicting that, on Election Day, Trump's vote share would be 46%. Under the simple (simplistic?) premise that undecided voters would ultimately split between Trump and Biden the same way as the decided ones (and assuming third-party candidates maintained their minimal projected vote-shares), the 46-46 split would be revised to 49%-49%. Making such adjustments, we present in Table 2 RCP-based projections, *FiveThirtyEight* projections, and actual election results, focusing on 14 states/districts where the election seemed close. (Those districts—one in Maine and one in Nebraska-- each have one Electoral College vote; for simplicity, we shall speak of 14 states.) These states constitute all those classified as toss-up by RCP and/or tipping states by FiveThirtyEight (hereafter *swing* *states*). It is in swing states where the outcome is not obvious (unlike California or Alabama) that polling accuracy is most important.

Table 2: Trump's Actual 2020 Vote Shares and His Projected Vote Shares in 14 Swing States, For FiveThirtyEight and Real Clear Politics (RCP)

	Actual Trump	Projected Trump Vote Share:	
State	Vote Share	Five Thirty Eight[a,b]	Real Clear Politics[c]
Arizona	49.06(%)	48.1 (-0.96)	47.77 (-1.29)
Georgia	49.24	49.2 (-.04)	48.93 (-0.31)
Florida	51.22	48.4 (-2.82)	49.10 (-2.12)
Iowa	53.08	50.0 (-3.08)	49.37 (-3.71)
Michigan	47.84	45.5 (-2.34)	46.54 (-1.30)
Minnesota	45.31	44.8 (-0.51)	45.53 (0.21)
Nevada	47.67	46.2 (-1.47)	47.23 (-0.44)
North Carolina	49.93	48.8 (-1.13)	48.71 (-1.22)
Ohio	53.27	49.8 (-3.47)	48.63 (-4.63)
Pennsylvania	48.84	47.3 (-1.54)	48.42 (-0.42)
Texas	52.06	50.3 (-1.76)	49.03 (-3.03)
Wisconsin	48.82	45.5 (-3.32)	45.08 (-3.74)
Maine, 2nd CD[d]	52.26	47.8 (-4.46)	46.56 (-5.70)
Nebraska, 2nd CD[d]	45.45	48.4 (2.95)	46.10. (0.65)
Average Forecast Error		-1.71	-1.93
Average Absolute Forecast Error		2.13	2.05

Notes:

a: Numbers in parenthesis are forecasting errors

b: Estimates by FiveThirtyEight are offered to the nearest tenth of a percent

c: For RCP, undecided voters in each survey were split between Trump and Biden the same way as voters who had decided. However, only about 2% of those canvassed were undecided in the local polls that RCP used for its final projections.

d: These congressional districts each elect one member of the Electoral College, and were the subject of forecasts by both FiveThirtyEight and RCP

As we see, *FiveThirtyEight* only marginally outperformed RCP. Like *FiveThirtyEight,* RCP underestimated Trump's vote shares in the swing states but RCP did so by slightly more; RCP, however, had the lower mean absolute forecast error.

Yet Table 2 does not contradict the possibility that individual polls in the various states performed badly. RCP, like *FiveThirtyEight,* is an aggregator of polls. Perhaps what happened is that large biases of some local polls were largely cancelled by opposite biases of other surveys. Even if that happened, though, the polls are not fatally flawed if the simple expedient of averaging their results yields a reasonably accurate outcome.

Actually, however, the local polls fared rather well in themselves. Table 3 presents for each state the average absolute forecast errors of all the local polls RCP used (i.e., not allowing for cancellations among opposite biases). (By definition, local forecast errors in each state had the same average as RCP, meaning that their overall average was -1.93 percentage points.). The absolute errors across the 14 entities averaged 2.53 percentage points, not much higher than *FiveThirtyEight*'s average of 2.13 percentage points. Table 3 also presents state-specific average margins of random sampling error for individual local polls, taking account of their sample sizes. These sampling margins generally exceeded absolute errors to an appreciable extent (on average, 3.83% vs. 2.53%).

Table 3: Average Absolute Forecast Error and Average Margin of Random Sampling Error for Trump's 2020 Vote Share, Among Key Local Polls in 14 Swing States[a]

State	Average Forecast Error[b]	Average Absolute Forecast Error[c]	Average Margin of Random Sampling Error[d]
Arizona	-1.29(%)	1.65	3.69
Florida	-2.12	2.72	3.54
Georgia	-0.31	1.49	4.00
Iowa	-3.71	3.71	3.82
Michigan	-1.30 (%)	2.10	3.96
Minnesota	0.21	0.54	3.41
Nevada	-0.44	1.50	3.45
N. Carolina	-1.22	1.25	3.67
Ohio	-4.63	4.63	3.33
Pennsylvania	-0.42	1.48	3.84
Texas	-3.03	3.03	3.39
Wisconsin	-3.74	3.74	3.63
Maine CD2	-5.70	5.70	5.59
Nebraska CD2	0.65	1.87	4.30
Average	-1.93	2.53	3.83

Notes:

a: Swing states are all those classified as toss-up by Real Clear Politics and/or tipping states by FiveThirtyEight

b: Includes all local polls used by Real Clear Politics in its final forecast of Trump's 2020 vote share

c: Arithmetical Average of Absolute Errors for the individual polls considered, which average five per state

d: Average Margins of Error for the individual polls considered, based on their sample sizes, estimated Trump and Biden vote shares, and estimated vote share for third-party candidates. Undecided voters were allocated to Trump and Biden in proportion to their vote shares among decided voters. The approximate formula for a given poll's margin of error (MOE) for Trump's vote share was:

$$MOE \approx 1.96 * \frac{q_T + q_B + q_U}{q_T + q_B} * \frac{\sqrt{q_T(1 - q_T)}}{\sqrt{n-1}}$$

Where:

$q_U = undecided\ vote\ share;\ q_T = Trump\ vote\ share;\ q_B = Biden\ \ vote\ share;\ n = sample\ size$

Against this backdrop, it is difficult to depict those local polls as major failures. Among the 70 local polls used by RCP across the 14 states, 77% yielded estimates of Trump's vote share within the usual confidence interval for that share based on sampling fluctuations alone. That figure does fall below the theoretical 95% level, meaning that it suggests the presence of some systematic error. But not huge systematic error.

Table 4 summarizes the comparison between FiveThirtyEight, RCP, and the local polls used by RCP in its final pre-election estimates.

Table 4: Average Forecasting Error and Average Absolute Forecasting Error for Trump's 2020 Vote Share in 14 Swing States[a], For *FiveThirtyEight*, Real Clear Politics, and Individual Local Polls[b]

	FiveThirtyEight	Real Clear Politics[c]	Local Polls
Average Error	-1.71(%)	-1.93	-1.93
Average Absolute Error	2.13	2.05	2.53

a: the 12 states and two congressional districts listed in Table Z

b: the local pre-election polls averaged by Real Clear Politics for its final estimate of Trump's vote share. There were an average of five such polls per state.

c: For Real Clear Politics and local polls, those undecided in local polls were allocated between Trump and Biden in proportion to their vote shares among decided voters. (Over the states considered, about two percent of voters were undecided.) The proportions of those canvassed who expressed support for third-party candidates were projected as the vote shares for those candidates on election day.

What about the other 42 states where the election was not considered close? RCP generally declined to offer forecasts there because reliable polls were scarce, presumably because the outcomes were viewed as foregone conclusions. But *FiveThirtyEight* did offer forecasts, although they often had to work with nonrandom polls like those from Survey

Monkey that the website itself had assigned the grade D-. Under the circumstances, *FiveThirtyEight* did well in the 42 noncompetitive states: Section 3.3 and Table 2 imply that its forecasts were too low on average by about two percentage points there, about the same as in the 14 swing states where serious polls were numerous.

Final Remarks

From the perspective of US public policy, a lesser role for presidential-election polls could have its advantages. Some of the energy spent obsessing over polls might be redirected to discussions of policy, while there could be lesser distortions like the bandwagon effect and abstentions from voting because the polls weren't close. Fewer voters might strategically decline to choose their preferred candidate. It is noteworthy in this connection that dozens of countries—including Canada, France, Greece, Mexico, Norway, and Poland---impose blackout periods on pre-election polls. The reasoning that led to those blackouts could well apply to the United States.

However, there is also a case that the primacy of polls in US elections for president, though not ideal, is better than the alternative. The dichotomy "polls versus policy issues" is overstated: The preferences that participants express in polls to a considerable extent reflect their views on policy matters. And when polls are close, they might stimulate voter turnout, reduce strategic voting, and induce candidates to speak at greater length on the virtues of their policy stances. There is also the deeper point that attempts to restrict polls could be viewed as antidemocratic in spirit, implying that voters should be denied information they want because they might use that information inappropriately. If voters cannot be trusted with polling results, should they likewise be deprived of facts on a variety of policy matters?

Yet all this discussion becomes moot if presidential polls are not viewed as trustworthy. We have considered such polls in 2020 in the swing states where accuracy was most important and where presidential candidates focused their campaigns. Whether working with *FiveThirtyEig*ht, Real Clear Politics, or the individual polls that offered "fuel" for such aggregators, we found performances in 2020 that were objectively very good. These performances were especially impressive given that the Covid-19 pandemic caused unprecedented difficulties in forecasting election results. The suggestion that the pollsters and aggregators failed in 2020 emerges as exaggerated, while the notion that biased individual polls required massive corrections from aggregators is inconsistent with relevant data.

Yet it is concerning that, for the second election in a row, the polls underestimated the support for Donald Trump and *FiveThirtyEight* did not devise an appropriate adjustment for the downward bias. Measures taken after the 2016 election to counteract the bias seem not to have fully succeeded, and the American Associate for Public Opinion Research (AAPOR) has explained at length that simple explanations for the problem do not readily fit the data. The only common explanation for the shortfall that AAPOR did not exclude was that Trump supporters may have refused to take part in voter surveys to a greater extent than Trump opponents, even within identifiable subgroups like white working-class voters or Republican voters. While one hopes that lessons from 2020 will avoid the problem in 2024, there is no certainty that this will be the case.

Those who think presidential polls get undue attention can continue to advance their arguments. But given what happened in 2020, those contending in 2024 that such polls should be ignored should not advance the assertion that the polls are highly unreliable.

References

Ad Hoc Committee on 2016 Election Polling (2017), "An Evaluation of 2016 Polls in the US, available at https://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx

S. Arnesen and O. Bergfjord (2014), "Prediction Markets vs. Polls: An Examination of Accuracy for the 2008 and 2012 Elections," *The Journal of Prediction Markets*, 8(3), pp. 24-32(2014)

J. Baker (2020), "Good Grief, the Pollsters Got It Wrong," *The Wall Street Journal*, November 10, 2020, available at https://www.wsj.com/articles/good-grief-the-pollsters-got-it-wrong-11605049069?mod=opinion_major_pos5

A.Barnett, (2018)"Epic Fail? The Polls and the 2016 Presidential Election, *Chance* 31(4) pp. 4-8

A. Blais, E. Gidengil, and N. Nevitte (2006), "Do Polls Influence the Vote?" pp. 264-279 in *Capturing Campaign Effects*, U. of Michigan Press

S. Bokat-Lindell (2020), "2016 Dealt a Blow to Polling. Did 2020 Kill It?" *The New York Times*, November 5, 2020, available at https://www.nytimes.com/2020/11/05/opinion/election-polls-wrong.html

B. Burden (2005), "Minor Parties and Strategic Voting in Recent U.S. Presidential Elections," *Electoral Studies*, 24(4), pp. 603-618, https://doi.org/10.1016/j.electstud.2005.02.004

L. Bursztyn, D. Cantoni, P. Funk, and N. Yuchtman (2017), "Polls, the Press, and Political Participation: The Effects of Anticipated Election Closeness on Voter Turnout, *CEPR Discussion Paper DP12088*, Available at SSRN: https://ssrn.com/abstract=2988846

W.J. Campbell (2020), "2020 Is Another Embarrassing Failure for Election Pollsters," *Fast Company*, November 4, 2020, available at https://www.fastcompany.com/90572224/2020-is-another-embarrassing-failure-for-election-pollsters

J. Clinton et al, (2021), "2020 Pre-Election Polling, An Evaluation of the 2020 General Election Polls," available at https://www.aapor.org/Education-Resources/Reports/2020-Pre-Election-Polling-An-Evaluation-of-the-202.aspx

M. Farjam (2021), "The Bandwagon Effect in an Online Voting Experiment With Real Political Organizations," *International Journal of Public Opinion Research*, 33(2) pp. 412-421, Summer 2021, https://doi.org/10.1093/ijpor/edaa008

Federal Election Commission (2021), *Official 2020 Presidential General Election Results*, available at https://www.fec.gov/resources/cms-content/documents/2020presgeresults.pdf

I. Ferre (2020), “The Biggest Election Takeaway Is the ‘Absolutely Massive Failure of Polling’: Tom Lee, *Yahoo! Finance*, November 4, 2020, available at https://www.yahoo.com/lifestyle/the-biggest-election-takeaway-is-the-absolutely-massive-failure-of-polling-tom-lee-205021837.html

FiveThirtyEight(2020)*, Final 2020 Presidential Election Forecast*, November 3, 2020, *FiveThirtyEight We*bsite, available at https://projects.fivethirtyeight.com/2020-election-forecast/

FiveThirtyEight (2021), “How Good Are FiveThirtyEight Election Forecasts?” *FiveThirtyEight Web*site, June 7, 2021, available at https://projects.fivethirtyeight.com/checking-our-work/

H. Olson (2020), “The Polling Industry Can’t Sweep Its Failure Under the Rug,” *The Washington Post*, November 4, 2020, available at https://www.washingtonpost.com/opinions/2020/11/04/henry-olsen-polling-industry-failure/

C. Panagopoulos (2021), “Polls and Elections: Accuracy and Bias in the 2020 US General Election Polls,” *Presidential Studies Quarterly*, 51(1), pp. 214-227 (March 2021)

C. Prosser and J. Mellon (2018), “The Twilight of the Polls? A Review of Trends in Polling Accuracy and the Causes of Polling Misses,” *Government and Opposition*, 53(4) pp. 757-790 (October 2018)

D. Rothschild (2009), “Forecasting Elections: Comparing Prediction Markets, Polls, and Their Biases,” *Public Opinion Quarterly*, 73(5), pp. 895-916

N. Silver (2020), “The Polls Weren’t Great. But That’s Pretty Normal,” *FiveThirtyEight Website*, November 11, 2020, available at https://fivethirtyeight.com/features/the-polls-werent-great-but-thats-pretty-normal/

S. Westwood, S. Messing, and Y. Lelkes (2020) , “Projecting Confidence: How the Probabilistic Horse Race Confuses and Demobilizes the Public**,”** *Journal of Politics*, 82(4), October 2020, https://doi.org/10.1086/708682

F. Wright and A. Wright (2018), “How Surprising Was Trump’s Victory? Evaluations of the 2016 US Presidential Election and a New Poll Aggregation Model, *Electoral Studies* 54, pp. 81-89

Appendix A: A Probability Distribution Based on *FiveThirtyEight* for the Number of Tipping States Trump Would Carry in 2020

We define S as the total number of the Tipping states Trump would carry (out of nine) and q_j as the probability he carries exactly j of those states. Then, based on *FiveThirtyEight* and analyses about its forecasts, we estimated in the main text that $E(S) = 1.82$, $and\ \sigma^2(S)\ \approx 1.64$; we also have $\sum_{j=0}^{9} q_j = 1.$ There are an infinite number of distributions for S that satisfy these three conditions but, to be practical, we followed the following procedure:

(i) To keep the number of feasible solutions finite, restrict the individual q_j 's to be multiples of .01 in the range 0 to 1.

(ii) Identify using an algorithm the combinations of q_j 's that match $E(S) = 1.82$, and $\sigma^2(S)\ \approx 1.64\ and\ \sum_{j=0}^{9} q_j = 1.$

(iii) Assign equal weight to all such combinations.

(iv) Average their values of q_j together to get a composite estimate of the probability that Trump would win exactly j of the swing states.

Steps (iii) and (iv) reflect the premise that, to get the best representation of what *FiveThirtyEight* implies about outcomes in the nine tipping states, it is reasonable to average over all distributions for S that are consistent with both the website's state-by-state probabilities and the assumption about cross-state correlation reflected in $\sigma^2(S)$. In a Bayesian sense, it is as if all probability distributions on the integers 0 to 9 were initially assumed equally likely to be correct (a uniform prior), and the distributions were updated by the requirements in (6) on mean and standard deviation. Those distributions that failed to meet the requirements were assigned a posterior probability of zero, while all the rest were assigned equal probabilities. Averaging q_j-values across the "surviving" distributions could therefore be viewed as yielding a reasonable expected value for q_j.

To implement (ii), we began by identifying $q_j's$ that matched $E(S) = 1.82 \; and \; \sum_{j=0}^{9} q_j = 1$, using an algorithm that rapidly excluded the overwhelming majority of $(q_0, q_1, \ldots q_9)$ combinations. For example, the procedure immediately excluded values of q_0 above 0.8, because, even if the remaining probability mass were assigned to q_9 , the mean would fall short of 1.82. For the same reason, q_9 could not exceed 0.20. Given a feasible value of q_0, the range of feasible values of q_1 could be identified. Continuing in this way, the algorithm generated the sets of values $(q_0, q_1, \ldots q_9)$ that yielded a mean of 1.82 (to the nearest hundredth).

For each set $(q_0, q_1, \ldots \; q_9)$ that yielded a mean of 1.82, the algorithm calculated the variance of the distribution, and retained only those sets with mean-squares of $1.82^2 + 1.64^2 \approx 4.95$ to the nearest hundredth. Then as described earlier, we created the "composite" distribution for S based on averaging the feasible distributions, which was:

$$S = \begin{cases} 0 \; w.p..094 \\ 1 \; w.p..328 \\ 2 \; w.p..380 \\ 3 \; w.p..118 \\ 4 \; w.p..041 \\ 5 \; w.p..018 \\ 6 \; w.p..010 \\ 7 \; w.p..005 \\ 8 \; w.p..004 \\ 9 \; w.p..002 \end{cases}$$

Appendix B: FiveThirtyEight's Own Model Validation Procedures

In a document at its website, *FiveThirtyEight* identifies its two procedures for validating its probabilistic forecasts against actual election results. The first one compares the probabilities assigned to events to the frequencies with which they actually occurred. For example, the website reports that "we'll throw every prediction …. between a 37.5 percent and 42.5 percent chance of winning into the same "40 percent" group — and then plot the averages of each bin's forecasted chances of winning against their actual win percentage." This is done for a full set of ranges to create a calibration plot and, it is stated, the data points should be "close to the 45 degree line" if the forecasts are all accurate.

A second evaluation method entails use of Brier skill scores. The initial Brier score BS for a set of n probabilistic forecasts follows:

$$BS = (\frac{1}{n}) \sum_{i=1}^{n} (p_i \; - \; O_i)^2$$

where p_i = estimated probability that event i will occur

$$O_i = \begin{cases} 1 \; if \; event \; i \; actually \; occurs \\ 0 \; if \; event \; i \; does \; not \; occur \end{cases}$$

BS is effectively a measure of mean-squared forecasting error and, the smaller is BS, the higher is the level of accuracy under this criterion.

The Brier Skill Score BSS compares BS to the Brier score BS_{ref} that would arise if a series of "unskilled" forecasts were advanced for the same n events. BSS takes the form:

$$BSS = 1 - BS/BS_{ref}$$

BSS is somewhat analogous to R^2 in linear regression analysis. *FiveThirtyEight* suggests that an unskilled forecast assumes that “each candidate has an equal shot.” In a two-candidate race, that would mean that the unskilled forecaster would assign win probabilities of ½ for both candidates. Then BS_{ref} would be ¼ because the quantity $(p_i - O_i)^2$ would necessarily be ¼ for all forecasts.

Comments on FiveThirtyEight’s Evaluation Methods

The website’s evaluation methods are reasonable in general, but they might be less so in the context of US presidential elections. The test with the 45-degree line works well if the various forecasts are independent, but can be misleading if the forecasts are correlated[i]. The Brier scores can also be problematic in the context of presidential elections. They reward predictions that are bold and accurate: if an event occurs, a 98% probability previously assigned to it is treated as far superior to a probability of 52%. The implicit premise is that the 98% reflects far greater insight than the equivocal 52%.

In the 2020 election, however, assigning strong probabilities to a Biden win in California or a Trump win in Wyoming was belaboring the obvious; the germane question is how well the forecaster performed in the “swing states” that would determine the winner. There a sophisticated forecaster processing all available information might sensibly assign a probability near 50% that Trump would win. Yet BSS would discount the forecaster’s skill in those states by likening their assessment to a coin toss. The issue is especially important because only about one-fifth of American states were swing states in 2020. Thus, the easy predictions in the other four-fifth of states will dominate the Brier score, potentially yielding a highly positive assessment of forecast accuracy even if the swing-state predictions fell short of the mark.

Importantly, FiveThirtyEight does not identify any procedures that mention correlation, or consider the accuracy of its vote-share distributions. Our tests that consider these issues, therefore, are not redundant.

[i] For example, suppose that a model correctly assumes that the candidate has a 50% chance of winning in each state, but that the various outcomes have strong positive correlation. Then the percentage actually won could be polarized towards 100% or 0%, and the unbiased estimates of 50% could appear highly inaccurate.