Towards Principled Unskewing: Viewing 2020 Election Polls Through a Corrective Lens From 2016

Michael Isakov¹ Shiro Kuriwaki^{2,3}

¹Harvard University, Cambridge, Massachusetts, United States of America,

²Department of Government, Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, United States of America,

³Department of Political Science, School of Humanities and Sciences, Stanford University, Stanford, California, United States of America

Published on: Nov 03, 2020

DOI: <https://doi.org/10.1162/99608f92.86a46f38>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

We apply the concept of the data defect index [Meng \(2018\)](#) to study the potential impact of systematic errors on the 2020 pre-election polls in 12 presidential battleground states. We investigate the impact under the hypothetical scenarios that (1) the magnitude of the underlying nonresponse bias correlated with supporting Donald Trump is similar to that of the 2016 polls, (2) the pollsters' ability to correct systematic errors via weighting has not improved significantly, and (3) turnout levels remain similar to 2016. Because survey weights are crucial for our investigations but are often not released, we adopt two approximate methods under different modeling assumptions. Under these scenarios, which may be far from reality, our models shift Trump's estimated two-party voteshare by a percentage point in his favor in the median battleground state, and increases twofold the uncertainty around the voteshare estimate.

Keywords: Bayesian modeling, survey nonresponse bias, election polling, data defect index, data defect correlation, weighting deficiency coefficient

Media Summary

Polling biases combined with overconfidence in polls led to general surprise in the outcome of the 2016 U.S. presidential election, and has resulted in an increased popular distrust in election polls. To avoid repeating such unpleasant polling surprises, we develop statistical models that translate the 2016 prediction errors into measures for quantifying data defects and pollsters' inability to correct them. We then use these measures to remap recent state polls about the 2020 election, assuming the scenarios that the 2016 data defects and pollsters' inability persist, and with similar levels of turnout as in 2016. This scenario analysis shifts the point estimates of the 2020 two-party voteshare by about 0.8 percentage points toward Trump in the median battleground state, and most importantly, increases twofold the margins of error in the voteshare estimates. Although our scenario analysis is hypothetical and hence should not be taken as predicting the actual electoral outcome for 2020, it demonstrates that incorporating historical lessons can substantially change—and affect our confidence in—conclusions drawn from the current polling results.

1. Polling Is Powerful, But Often Biased

Polling is a common and powerful method for understanding the state of an electoral campaign, and in countless other situations where we want to learn some characteristics about an entire population but can only afford to seek answers from a small sample of the population. However, the magic of reliably learning about many from a very few depends on the crucial assumption of the sample being representative of the population. In layperson's terms, the sample needs to look like a miniature of the population. Probabilistic sampling is the only method with rigorous theoretical underpinnings to ensure this assumption holds [Meng \(2018\)](#). Its

advantages have also been demonstrated empirically, and is hence the basis for most sample surveys in practice ([Yeager et al., 2011](#)). Indeed, even nonprobabilistic surveys use weights to mimic probabilistic samples ([Kish, 1965](#)).

Unfortunately, representative sampling is typically destroyed by selection bias, which is rarely harmless. The polls for the 2016 U.S. presidential election provide a vivid reminder of this reality. Trump won the toss-up states of Florida and North Carolina but also states like Michigan, Wisconsin, and Pennsylvania, against the projections of most polls. Poll aggregators erred similarly; in fact, [Wright and Wright \(2018\)](#) suggest that overconfidence was more due to aggregation rather than the individual polling errors. In the midterm election 2 years later, the direction and magnitude of errors in polls of a given state were heavily correlated with their own error in 2016 ([Cohn, 2018](#)). Systematic polling errors were most publicized in the 2016 U.S. presidential election, but comparative studies across time and countries show that they exist in almost any context ([Jennings & Wlezien, 2018](#); [Kennedy et al., 2018](#)) [Shirani-Mehr et al. \(2018\)](#).

In this article we demonstrate how to incorporate historical knowledge of systematic errors into a probabilistic model for poll aggregations. We construct a probabilistic model to capture selection mechanisms in a set of 2016 polls, and then propagate the mechanisms to induce similar patterns of hypothetical errors in 2020 polls. Our approach is built upon the concept of the data defect index [Meng \(2018\)](#), a quantification of nonrepresentativeness, as well as the postelection push for commentators to ‘unskew’ polls in various ways. The term ‘unskew’ is used colloquially by some pollsters, loosely indicating a correction for a poll that overrepresents a given party’s supporters. Our method moves toward a statistically principled way of unskewing by using past polling error as a starting point by decomposing the overall survey error into several interpretable components. The most critical term of this identity is an index of data quality, which captures total survey error not taken into account in the pollster’s estimate.

Put generally, we provide a framework to answer the question, ‘What do the polls imply if they were as wrong as they were in a previous election?’ If one only cared about the point estimate of a single population, one could simply subtract off the previous error. Instead, we model a more complete error structure as a function of both state-specific and pollster-specific errors through a multilevel regression, and properly propagate the increase in uncertainty as well as the point estimate through a fully Bayesian model.

Our approach is deliberately simple, because we address a much narrower question than predicting the outcome of the 2020 election. We focus on unskewing individual survey estimates, not unskewing the entire forecasting machinery. Therefore our model must ignore non-poll data, such as economic conditions and historical election results, which we know would be predictive of the outcome ([Hummel & Rothschild, 2014](#)). Our model is also simplified (restricted) by using only point estimates reported by pollsters instead of their microdata, which generally are not made public. Since most readers focus on a state’s winner, we model the voteshare only as a proportion of the two-party’s votes, implicitly netting out third-party votes or undecided voters. We largely avoid the complex issue of overtime opinion change over the course of the campaign (e.g.,

[Linzer, 2013](#)), although we show that our conclusions do not change even with modeling a much shorter time period where opinion change is unlikely. We do not explicitly model the correlation of prediction errors across states, which would be critically important for reliably predicting any final electoral college outcome.

Finally, the validity of our unskewing is only as good as the hypothetical we posit, which is that the error structure has not changed between 2016 and 2020. For example, several pollsters have incorporated lessons from 2016 ([Kennedy et al., 2018](#)) in their new weighting, such as adding education as a factor in their weighting ([Skelley & Rakich, 2020](#)). Such improvement may likely render our ‘no-change’ hypothetical scenario overly cautious. On the other hand, it is widely believed that 2020 turnout would be significantly higher than in 2016. Because polling biases are proportionally magnified by the square root of the size of the turnout electorate [Meng.\(2018\)](#), our use of 2016 turnout in place for 2020 may lead to systematic undercorrections. Whereas we do not know which of these factors has stronger influence on our modeling (a topic for further research), our approach provides a principled way of examining the current polling results (i.e., the now-cast), with a confidence that properly reflects the structure of historical polling biases.

The rest of the article is organized as follows. First, we describe two possible approaches to decompose the actual polling error into more interpretable and fundamental quantities. Next, we introduce key assumptions that allow us to turn these decompositions into counterfactual models, given that polling bias (to be precisely defined) in 2020 is similar to that in the previous elections. Finally, we present our findings for 12 key battleground states, and conclude with a brief cautionary note.

2. Capturing Historical Systematic Errors

Our approach uses historical data to assess systematic errors, and then use them as scenarios—that is, what if the errors persist—to investigate how they would impact the current polling results. There are at least two kinds of issues we need to consider: (1) defects in the data; and (2) our failure to correct them. Below we start by using the concept of “data defect correlation” of [Meng.\(2018\)](#) to capture these considerations, and then propose a (less ideal) variation to address a practical issue that survey weights are generally not reported.

2.1. Data Defect Correlation

Whereas it is clearly impossible to estimate the bias of a poll from itself, the distortion caused by nonsampling error is a modelable quantity. [Meng.\(2018\)](#) proposed a general framework for quantifying this distortion by decomposing the actual error into three determining factors: data quality, data quantity, and problem difficulty. Let I index a finite population with N individuals, and Y_I be the survey variable of interest (e.g., $Y_I = 1$ if the I th individual plans to vote for Trump, and $Y_I = 0$ otherwise). We emphasize that here the random variable is not Y , but I , because the sampling errors are caused by how opinions vary among individuals as opposed to uncertainties in individual opinions Y_I for fixed I (at the time of survey). Following the standard finite population probability calculation ([Kish, 1965](#)), the population average $\bar{Y}_N$ then can be written as the

mean $E(Y_I)$, where the expectation is with respect to the discrete uniform distribution of I over the integers $\{1, \dots, N\}$.

To quantify the data quality, we introduce the data recording indicator R_I , which is one if the I th individual's value Y_I is recorded (i.e., observed) and zero otherwise. Clearly, the sample size (of the observed data) then is given by $n = \sum_{\ell=1}^N R_\ell$. Here R_I captures the overall mechanism that resulted in individual answers being recorded. If everyone responds to a survey (honestly) whenever they are selected, then R merely reflects the sampling mechanism. However, in reality there are almost always some nonrespondents, in which case R captures both the sampling and response mechanisms. For example, $R_I = 0$ can mean that the I th individual was not selected by the survey or was selected but did not respond. To adjust for nonresponse biases and other imperfections in the sampling mechanism, survey pollsters typically apply a weighting adjustment to redistribute the importance of each observed Y to form an estimator, say, for the population mean $\bar{Y}_N$ in the form of

$$(2.1) \quad \bar{Y}_{n,w} = \frac{\sum_{\ell=1}^N w_\ell R_\ell Y_\ell}{\sum_{\ell=1}^N w_\ell R_\ell},$$

where w_ℓ indicates a calibration weight for respondent ℓ to correct for known discrepancies from the target population by the observed sample. Typically weights are standardized to have mean 1 such that the denominator in (2.1) is n , a convention that we follow.

Let $\rho = \text{Corr}(w_I^*, Y_I)$, the population (Pearson) correlation between $w_I^* \equiv w_I R_I$ and Y_I (again, the correlation is with respect to the uniform distribution of I), and σ be the population standard deviation of Y_I . Define n_w to be the *effective sample size* induced by the weights ([Kish, 1965](#)), that is,

$$n_w = \frac{\sum_{\ell=1}^N (R_\ell w_\ell)^2}{\sum_{\ell=1}^N R_\ell w_\ell^2} = \frac{n}{1 + s_w^2},$$

where s_w^2 is the sample—not population—variance of the weights (but with divisor n instead of $n - 1$). It is then shown in [Meng \(2018\)](#) that the actual estimation error

$$(2.2) \quad e_Y \equiv \bar{Y}_{n,w} - \bar{Y}_N = \rho \cdot \sqrt{\frac{1 - f_w}{f_w}} \cdot \sigma,$$

where $f_w = n_w/N$ is the *effective sampling fraction*. Identity (2.2) tells us that (i) the larger the ρ in magnitude, the larger the estimation error, and (ii) the direction of error is determined by the sign of ρ . Probabilistic sampling and weighting ensures ρ is zero on average, with respect to the randomness introduced by R . But this is no longer the case when (a) R_I is influenced by Y_I because of individuals' selective response behavior or (b) the weighting scheme fails to fully correct this selection bias. Hence, ρ measures the ultimate defect in the data due to (a) and (b) manifesting itself in the estimator $\bar{Y}_{n,w}$. Consequently, it is termed a *data defect correlation* (DDC).

There are two main advantages of using ρ to model data defect instead of directly shifting the current polls by the historical actual error e_Y (e.g., from 2016 polls). First, it disentangles the more difficult systematic error from the well-understood probabilistic sampling errors due to the sample size (encoded in f_w) and the problem difficulty (captured by σ). Using e_Y for correcting biases therefore would be mixing apples and oranges, especially for polls with varying sizes. This can be seen more clearly by noting that (2.2) implies that the mean-squared error (MSE) of $\bar{Y}_{n,w}$ can be written as

$$E_R(\bar{Y}_{n,w} - \bar{Y}_N)^2 = N \cdot E_R(\rho^2) \left[(1 - f_w) \frac{\sigma^2}{n_w} \right],$$

where the expectation is with respect to the report indicator R (which includes the sampling mechanism). We note that the term in the brackets is simply the familiar variance of sample mean under simple random sampling (SRS) with sample size n_w ([Kish, 1965](#)). It is apparent then that the quantity $E_R(\rho^2)$, termed *data defect index (ddi)* in [Meng \(2018\)](#), captures any increase (or decrease) in MSE beyond SRS *per individual in the population* (because of the multiplier N).

Second, in the case where the weights are employed only for ensuring equal probability sampling, ρ measures the individual response behavior, which is a more consequential and potentially more stable measure over time. The fact that ddi captures the design gain or defect per individual in population reflects its appropriateness for measuring individual behavior. This observation is particularly important for predictive modeling, where using quantities that vary less from election to election can substantively reduce predictive errors.

2.2. Weighting Deficiency Coefficient

Pollsters create weights through calibration to known population quantities of demographics, aiming to reduce sampling or nonresponse bias especially for nonprobability surveys (e.g., [Caughey et al., 2020](#)). The effective sample size n_w is a compact statistic to capture the impact of weighting in the DDC identity (2.2), but is often not reported. If n_w does not vary (much) across polls, then its impact can be approximately absorbed by a constant term in our modeling, as we detail in the next section. However, this simplification can fail badly when n_w varies substantially between surveys, particularly in ways that are not independent of ρ .

To circumvent this problem, we consider an alternative decomposition of e_Y by using the familiar relationship between correlation and regression coefficient, which leads to

$$(2.3) \quad e_Y = \eta \cdot f^{-1} \cdot \sigma^2,$$

where $f = n/N$. Here η is the population regression coefficient from regressing w_I^* on Y_I , that is,

$$\eta = \rho \frac{\sqrt{\text{Var}(w_I^*)}}{\sigma}.$$

Identity (2.3) then follows (2.2) and that

$$\text{Var}(w_I^*) = (1 + s_w^2)^2 f_w (1 - f_w) = f(1 - f + s_w^2).$$

Being a regression coefficient, the term η governs how much variability in w_I^* can be explained by Y_I and thus it expresses the association between the variable of interest and a weighted response indicator. Since the weights are intended to reduce this association, it can be interpreted as how efficiently the survey weights correct for a biased sampling mechanism. This motivates terming η as the *weighting deficiency coefficient* (WDC): the larger it is in magnitude, the less successful the weighting scheme is in reducing the data defect caused by the selection mechanism.

Therefore, if we assume our ability in using weights to correct bias, captured by the regression coefficient η , has not been substantially improved since 2016, then we can bypass the need of knowing n_w for either election when we conduct a scenario analysis. Importantly, we must recognize that this is different from assuming ρ remains the same. Since fixing either ρ or η is a new way of borrowing information from the past, we currently do not have empirical evidence to support which one is more realistic in practice. However, the fact that fixing η bypasses the need of knowing the effective sample size makes it practically more appealing when this quantity is not available.

We recognize that n_w can be recovered from a pollster's methodology report when the reported margin of error (MoE) of a poll estimator $\hat{\mu}$ is based on the standard formula that accounts for the design effect through the effective sample size n_w : $\text{MoE} = 2\sqrt{\hat{\mu}(1 - \hat{\mu})/n_w}$. In that case, $n_w = 4\hat{\mu}(1 - \hat{\mu})/\text{MoE}^2$. We therefore strongly recommend pollsters to either report how the MoE is computed or the effective sample size due to weighting.

2.3. Regression Framework With Multiple Polls

Whereas we cannot estimate ρ or η from a single poll, identities (2.2) and (2.3) suggest that we can treat them as regression coefficients from regressing $\bar{Y}_{n,w}$ on X , where $X = \sqrt{(1 - f_w)f_w^{-1}}\sigma$ and $X = f^{-1}\sigma^2$ respectively, when we have multiple polls. That is, the variations of f_w or s_w^2 in polls provide a set of design points X_k , where k indexes different polls, for us to extract information about ρ or η , if we believe ρ (or η) is the same across different polls. By invoking a Bayesian hierarchical model, we can also permit ρ 's or η 's to vary with polls, with the degree of similarity between them controlled by a prior distribution.

As we pointed out earlier, a current practical issue for using ρ is that pollsters do not report s_w^2 (or n_w). In addition to switching to work with η , if s_w^2 's do not vary much with the polls, then we can treat $1 - f + s_w^2$ as a constant when $f \ll 1$, which typically is the case in practice. This leads to replacing X by $\hat{X} = f^{-1/2}\sigma$, as an approximate method (which we will denote by DDC*) for working with ρ , since a constant multiplier of the X can be absorbed by the regression coefficient. In contrast, when working with η , we can set X exactly as $X = f^{-1}\sigma^2$, which is the square of the approximate $\hat{X}$ for ρ .

For applications to predicting elections, especially those where the results are likely to be close, we can make a further simplification to replace $\sigma = \sqrt{\bar{Y}_N(1 - \bar{Y}_N)}$ by its upper bound 0.5, which is achieved when $\bar{Y}_N = 0.5$. The well-known stability of $p(1 - p) \approx 1/4$ when $0.35 \leq p \leq 0.65$ makes this approximation applicable for most polling situations where predictive modeling is of interest.

Accepting these approximations, the only critical quantity in setting up the design points is $f = n/N$. This is in contrast to traditional perspectives of probabilistic sampling, where what matters is the absolute sample size n , not the relative sampling rate $f = n/N$ (more precisely the fraction of the observed response). The fundamental reason for this change is precisely the identity (2.2), which tells us that as soon as we deviate from probabilistic sampling, the population size N will play a critical role in the actual error.

When we deal with one population, N is a constant for different polls, and hence it can be absorbed into the regression coefficient. However, to increase the information in assessing DDC or WDC, we will model multiple populations (e.g., states) simultaneously. Therefore, it is important to determine what N represents, and in particular, to consider “What is the population in question?” This is a challenging question for election polls because we seek to estimate the sentiment of people who *will vote*, but this population is unknowable prior to election day (Cohn, 2018; (Rentsch et al., 2019)). Pollsters use likely voter models to incorporate them in their sampling frames, and others use the population of registered voters. But these models are predictions: for example, not all registered voters vote, and dozens of state allow for same day registration.

Because total turnout is unknown before the election, we use the 2016 votes cast for the president in place of this quantity: $\hat{N} = N^{\text{pre}}$. A violation to this assumption in the form of larger turnout in 2020 could affect our results, because our regression term includes N for both DDC* and WDC and we are putting priors on the bias in 2020. In the past six presidential elections, a state’s voter turnout (again, as a total count of votes rather than a fraction) sometimes increased as much as 20%. Early indicators suggest a surge in 2020 turnout (McCarthy, 2019). With higher turnout, we would rescale our change in our point estimates of voteshare by approximately $\sqrt{N/N^{\text{pre}}}$ for DDC* and N/N^{pre} for WDC models, and further increase uncertainty. The problem of predicting voter turnout is important yet notoriously difficult (Rentsch et al., 2019), so incorporating this into our approach is a major area of future work. More generally, this underscores once again the importance of assumptions in interpreting our scenario analyses.

3. Building a Corrective Lens From the 2016 Election

To perform the scenario analysis as outlined in Section 1, we apply the regression framework to both 2016 and 2020 polling data. We use the 2016 results to build priors to inform and constrain the analysis for 2020 by reflecting the scenarios we want to investigate.

For any given election (2016 or 2020), our data include a set of poll estimates of Trump’s two-party voteshares (against Clinton in 2016 and Biden in 2020). Because we use data from multiple states (indexed by i) and multiple pollsters (indexed by j), we will denote each poll’s weighted estimate as y_{ijk} (instead of $\bar{Y}_{n,w}$) for

polls $k = 0, 1, \dots, K_{ij}$, where K_{ij} is the total number of polls conducted by pollster j in state i . We allow K_{ij} to be zero, because not every pollster conducts polls in every state. Accordingly, we set $X_{ijk} = f_{ijk}^{-1/2}$ (for DDC*) or $X_{ijk} = f_{ijk}^{-1}$ (for WDC), where $f_{ijk} = n_{ijk}/N_i$. (For simplicity of notation, we drop the subscript w here since all y_{ijk} 's use weights.)

3.1. Data

Various news, academic, and commercial pollsters run polls, make public their weighted estimates, if not their individual respondent-level data. For each poll, we consider its reported two-party estimate for Trump and sample size. While some pollsters (such as Marquette) focus on a single state, others (such as YouGov) poll multiple states. We focus on 12 battleground states and 17 pollsters that polled in those states in both 2016 and 2020. When a poll reports both an estimate weighted to registered voters and to likely voters, we use the former. We downloaded these topline estimates from FiveThirtyEight, which maintains a live feed of public polls.

Table 1. State polls used

(A) By State

	2020		2016	
State	Polls	Pollsters	Polls	Pollsters
Pennsylvania	22	10	54	11
North Carolina	18	10	61	13
Wisconsin	18	8	29	7
Arizona	17	9	32	9
Florida	14	10	58	13
Michigan	14	8	33	7
Georgia	11	7	27	8
Texas	10	4	22	6
Iowa	8	5	23	7
Minnesota	7	6	18	4

Ohio	6	4	50	11
New Hampshire	5	4	51	10
Total	150	17	458	17

(B) By Pollster

	2020		2016	
Pollster	Polls	States	Polls	States
Ipsos	24	6	165	12
YouGov	23	11	35	12
New York Times / Siena	18	12	6	3
Emerson College	13	10	28	11
Quinnipiac University	12	6	23	6
Monmouth University	11	6	16	9
Public Policy Polling	11	6	26	10
Rasmussen Reports	8	6	57	5
Suffolk University	6	6	9	6
SurveyUSA	6	3	8	5
CNN / SSRS	4	4	10	5
Marist College	4	4	15	8
Data Orbital	3	1	7	1
Marquette University	3	1	4	1

University of New Hampshire	2	1	11	1
Gravis Marketing	1	1	25	10
Mitchell Research & Communications	1	1	13	1

Note. 2020 polls are those taken from August 3 to October 21, 2020. 2016 polls are those taken from August 6 to November 6, 2016. Poll statistics are taken from FiveThirtyEight.com.

Table 1 summarizes the contours of our data. After subsetting to 17 major pollsters and limiting to pollsters with a FiveThirtyEight grade of C or above, we are left with 458 polls conducted in the 3 months leading up to the 2016 election (November 8, 2016), as well as 150 polls in 2020 taken during August 3, 2020, to October 21, 2020. Each of these battleground states were polled by two to 12 unique pollsters, totaling four to 21 polls. Each state poll canvasses about 500 to 1,000 respondents, and the total number of 2020 respondents in a given state range from 5,233 respondents in Iowa to 11,295 in Michigan.

Hindsight should allow us to calculate the DDC and WDC from 2016 (ρ^{pre} and η^{pre} , respectively) because we know the value of e_Y . However, because s_w needed by (2.2) is not available, as we discussed before, we set it to zero to obtain an upper bound in magnitude for DDC, which we will denote by DDC*. This allows us to use the ddi R package (Kuriwaki, 2020) directly. We compute past WDC of (2.3) exactly because it was designed to bypass the need for s_w . Table 2 provides the simple averages of the actual errors, DDC* and WDC from the 2016 polls we use.

Differences by pollster, shown in Table 2, Panel B, can arise from variation in pollster’s mode of the survey, sampling methods, and weighting methods. These differences can manifest themselves in systematic over- or underestimates of a candidate’s share—so called ‘house effects.’ Thus, we consider these effects together when studying polling bias.

A simple inspection indicates that the average WDCs vary nontrivially by pollsters, although all of them underestimated Trump’s voteshares by states (individual results not shown). An F test on the 458 values of WDC by state or pollster groupings rejects the null hypothesis that state-specific or pollster-specific means are the same ($p < 0.001$). In a two-way ANOVA, state averages comprise 41% of the total variation in WDC and pollster averages comprise 16%.

It is worth remarking that the values of DDC* in Table 2 are small (and hence DDCs are even smaller), less than a 10th of a *percentage point* in total. Meng (2018) found that the unweighted DDCs for Trump’s vote centered at about -0.005 , or half a percentage point, in the Cooperative Congressional Election Study. The values here are much smaller because they use estimates that already weight for known discrepancies in sample

demographics. Therefore, the resulting weighted DDC measures the data defects that remain after such calibrations.

3.2. Modeling Strategy

As we outlined in Section 2.3, our general approach is to use the 2016 data on DDC, or more precisely, DDC*, shown in Table 2 to model the structure of the error, and apply it to our 2020 polls via casting (2.2) as a regression. Similarly, we rely on (2.3) to use the structure of WDC (η) to infer the corrected (true) population average. Our Bayesian approach also provides a principled assessment of the uncertainty in our unskewing process itself.

Table 2. Average data defect correlation for Donald Trump in 2016.

(A) By State

State	Actual	Error	DDC*	WDC	Polls
Ohio	54.3%	-4.6 pp	-0.00096	-0.000024	50
Iowa	55.1%	-4.1 pp	-0.00144	-0.000052	23
Minnesota	49.2%	-3.8 pp	-0.00112	-0.000027	18
New Hampshire	49.8%	-3.2 pp	-0.00163	-0.000080	51
North Carolina	51.9%	-3.2 pp	-0.00073	-0.000018	61
Wisconsin	50.4%	-3.0 pp	-0.00085	-0.000028	29
Pennsylvania	50.4%	-2.9 pp	0.00058	-0.000014	54
Michigan	50.1%	-2.8 pp	0.00078	-0.000022	33
Florida	50.6%	-1.6 pp	0.00028	-0.000006	58
Arizona	51.9%	0.2 pp	0.00004	0.000003	32
Georgia	52.7%	0.5 pp	0.00007	0.000002	27
Texas	54.7%	2.6 pp	0.00041	0.000004	22
Total		-2.4 pp	0.00063	-0.000023	458

(B) By Pollster

Pollster	Error	DDC*	WDC	Polls
University of New Hampshire	-4.9 pp	-0.0023	-0.000135	11
Mitchell Research & Communications	-4.1 pp	-0.0010	-0.000036	13
Monmouth University	-3.6 pp	-0.0007	-0.000018	16
Marquette University	-3.3 pp	-0.0011	-0.000036	4
Public Policy Polling	-3.3 pp	-0.0009	-0.000044	26
YouGov	-3.2 pp	-0.0011	-0.000044	35
Marist College	-3.0 pp	-0.0006	-0.000028	15
Rasmussen Reports	-2.9 pp	-0.0007	-0.000022	57
Quinnipiac University	-2.7 pp	-0.0005	-0.000019	23
New York Times / Siena	-2.3 pp	-0.0005	-0.000013	6
Emerson College	-2.1 pp	-0.0008	-0.000024	28
Suffolk University	-2.1 pp	-0.0005	-0.000012	9
SurveyUSA	-2.0 pp	-0.0006	-0.000013	8
Ipsos	-1.8 pp	-0.0004	-0.000012	165
Gravis Marketing	-1.7 pp	-0.0004	-0.000013	25
CNN / SSRS	-0.9 pp	-0.0003	-0.000003	10
Data Orbital	-0.4 pp	-0.0002	-0.000003	7

Note. The “Actual” column in the first panel is Trump’s actual two-party voteshare in 2016. “Error” is the simple average of the poll predictions minus Trump’s actual voteshare, so that a negative value indicates *underestimation* of Trump. DDC* is the upper bound

of the data defect correlation ρ obtained from (2.2) but by setting $n_w = n$ (i.e., assuming $s_w = 0$), WDC is the weighting deficiency coefficient η of (2.3). All values are simple averages across polls, hence they do not rule out the possibilities for the average value of DDC* or WDC to have a different sign from the sign of the average actual error.

Specifically, let μ_i be a candidate's (unobserved) two-party voteshare in state i , and μ_i^{pre} the observed voteshare for the same candidate (or party) in a previous election. In our example, μ_i^{pre} denotes Trump's 2016 voteshare, listed in Table 2 (Panel A). A thorny issue regarding the current μ_i is that voters' opinions change over the course of the campaign ([Gelman & King, 1993](#)), though genuine opinion changes tend to be rare ([Gelman et al., 2016](#)). Whereas models for opinion changes do exist (e.g., [Linzer, 2013](#)), this scenario analysis amounts to contemplating a 'time-averaging scenario,' which can still be adequate for the purposes of examining the impact of historical lessons. However, we still reexamine our data with a much shorter timespan of 3 weeks where underlying opinion change is much less likely ([Shirani-Mehr et al. \(2018\)](#)). In any case, incorporating temporal variations can only increase our uncertainty above and beyond the results we present here.

We know from the study of uniform swing in political science that a state's voteshare in the past election is highly informative of the next election, especially in the modern era ([Jackman, 2014](#)). In every presidential election since 1996, a state's two-party voteshare was correlated with the previous election at a correlation value 90% or higher. This motivates us to model a state's *swing*: $\delta_i = \mu_i - \mu_i^{\text{pre}}$, instead of μ_i directly, for better control of residual errors. A single poll k by pollster j constitutes an estimate for the swing: $\hat{\delta}_{ijk} = y_{ijk} - \mu_i^{\text{pre}}$, where, as noted before, y_{ijk} is the weighed estimate from the k th poll. Then (2.3) can be thought of as a regression:

$$(3.1) \quad \begin{aligned} \hat{\delta}_{ijk} &= y_{ijk} - \mu_i + \mu_i - \mu_i^{\text{pre}} \\ &= \eta_{ijk} \cdot f_{ijk}^{-1} \cdot \sigma_i^2 + \delta_i \equiv \eta_{ijk} X_{ijk} + \delta_i, \end{aligned}$$

where η_{ijk} and f_{ijk} are respectively the realizations of η and f_w in (2.3) for y_{ijk} . Identity (3.1) then provides a regression-like setup for us to estimate δ_i as the intercept, when we can put a prior on η_{ijk} . For our scenario analysis, we will first use 2016 data to fit a (posterior) distribution for η_{ijk}^{pre} , which will be used as the prior for the 2020 η_{ijk} . We then use this prior together with the 2020 individual poll estimates y_{ijk} to arrive at a posterior distribution for δ_i and hence for μ_i .

Our modeling strategies for ρ and η are the same, with the only difference being that we use a Beta distribution for ρ (because it is bounded) but the usual Normal model for η . We will present the more familiar Normal model for WDC in the main text, and relegate the technical details of the Beta regression to the Appendix.

3.3. Assuming No Selection Bias

As a simple baseline for an aggregator, we formulate the posterior distribution of the state's voteshare when $E(\eta_{ijk}) = 0$, that is, when the only fluctuation in the polls is sampling variability. In this case, (3.1) implies

that δ_{ijk} is centered around δ_i , and its variability is determined by its sample size n_{ijk} . Typically we would also model its distribution by a Normal distribution, but to account for potential other uncertainties (e.g., the uncertainties in weights themselves), we adopt a more robust Laplace model, with scale $1/2$ to match the $1/4$ upper bound on the variance of a binary outcome. This leads to our baseline (Bayesian) model

$$(3.2) \quad \begin{aligned} \sqrt{n_{ijk}}(\hat{\delta}_{ijk} - \delta_i) &\stackrel{\text{i.i.d.}}{\sim} \text{Laplace}\left(0, \frac{1}{2}\right), \\ \delta_i | \mu_\delta, \tau_\delta^2 &\stackrel{\text{i.i.d.}}{\sim} N(\mu_\delta, \tau_\delta^2), \end{aligned}$$

where μ_δ can be viewed as the swing at the national level. By using historical data, we found the priors $\tau_\delta \sim 0.1 \cdot \log \text{Normal}(-1.2, 0.3)$ and $\mu_\delta \sim N(0, 0.03^2)$ are reasonably weakly informative (e.g., it suggests the rarity for the swing to exceed 6 percentage points, which indeed has not happened since 1980; the average absolute swing during 2000-2016 was 2.7 percentage points).

The Bayesian model will lead naturally to shrinkage estimation on cross-state differences in swings (from 2016 election to 2020 election). On top of being a catch-all for possible deviations from the simple random sample framework, the Laplace model has the interpretation of using a median regression instead of mean regression ([Yu & Moyeed, 2001](#)).

3.4. Modeling the WDC of 2016

In order to show how historical biases can affect the results from the above (overconfident) model, we assume a parametric form for η_{ijk} of polls conducted in 2016 and then incorporate this into (3.3). Given the patterns in Table 2, we model η_{ijk} using a multilevel regression with state and pollster-level effects for the mean and variance. We pool the variances on the log scale to transform the parameter to be unbounded. Specifically, we model:

$$(3.3) \quad \begin{aligned} \eta_{ijk}^{\text{pre}} | \theta_{ij}, \sigma_{ij} &\sim N(\theta_{ij}, \sigma_{ij}^2), \\ \theta_{ij} &= \gamma_0 + \gamma_i^{\text{state}} + \gamma_j^{\text{pollster}}, \\ \sigma_{ij} &= \exp(\phi_0 + \phi_i^{\text{state}} + \phi_j^{\text{pollster}}), \end{aligned}$$

where (the generic) $\gamma^{\text{state}}, \gamma^{\text{pollster}}, \phi^{\text{state}}, \phi^{\text{pollster}}$ are four normal random effects. That is, $\{\gamma_i^{\text{state}}, i = 1, \dots, I\} \stackrel{\text{i.i.d.}}{\sim} N(0, \tau_\gamma^2)$, and similarly for the other three random effects. Each of the four prior variances $\tau_\star^2$ is itself given a weakly informative boundary-avoiding $\text{Gamma}(5, 20)$ hyper-prior. The two intercepts γ_0 and ϕ_0 are given an (improper) constant prior.

The θ_{ij} can be interpreted as the mean DDC for pollster j conducting surveys in state i , while the σ_{ij} determines the variations in DDC in that state-pollster pair. The shrinkage effect induced by (3.3), on top of being computationally convenient, attempts to crudely capture correlations between states and pollsters when considering systematic biases, as was the case in the 2016 election.

3.5. Incorporating 2016 Lessons for 2020 Scenario Modeling

By (3.1), the setup in (3.3) naturally implies a model for estimating swing from 2016 to 2020:

$$(3.4) \quad \begin{aligned} \hat{\delta}_{ijk} | \delta_i, \theta_{ij}, \sigma_{ij} &\sim N(\delta_i + X_{ijk} \theta_{ij}, (X_{ijk} \sigma_{ij})^2), \\ \theta_{ij} &= \gamma_0 + \gamma_i^{\text{state}} + \gamma_j^{\text{pollster}}, \\ \sigma_{ij} &= \exp(\phi_0 + \phi_i^{\text{state}} + \phi_j^{\text{pollster}}), \end{aligned}$$

Although (3.4) mirrors the model for 2016 data, namely (3.3), there are two major differences. First, the main quantity of interest is now δ_i , not η_{ijk} . Second, we replace the normal random effect models for (3.3) by informative priors derived from the posteriors we obtained using the 2016 data. Specifically, let γ_v be any member of the collection of the γ -parameters, denoted by $\{\gamma_0, \gamma_i^{\text{state}}, \gamma_j^{\text{pollster}}, i, j = 1, \dots\}$ and similarly let ϕ_v be any component of the ϕ -parameters: $\{\phi_0, \phi_i^{\text{state}}, \phi_j^{\text{pollster}}, i, j = 1, \dots\}$. Then, for computational efficiency, we use the normal approximations as emulators to their actual posteriors derived from (3.3), that is, we assume

$$(3.5) \quad \begin{aligned} \gamma_v &\sim N(E_{\text{mc}}(\gamma_v^{\text{pre}}), \text{Var}_{\text{mc}}(\gamma_v^{\text{pre}})), \\ \phi_v &\sim N(E_{\text{mc}}(\phi_v^{\text{pre}}), \text{Var}_{\text{mc}}(\phi_v^{\text{pre}})), \end{aligned}$$

where the “pre” superscripts denote previous election, and E_{mc} and Var_{mc} indicate respectively the posterior mean and posterior variance obtained via MCMC (in this case with 10,000 draws). We use the same weakly informative priors on $\delta_i \stackrel{\text{i.i.d.}}{\sim} N(\mu_\delta, \tau_\delta^2)$ as in the last line of (3.2) to complete the fully Bayesian specification.

In using the posterior from 2016 as our informative prior for the current election, we have opted to put priors on the random effects as opposed to the actual θ_{ij} or σ_{ij} . This has the appealing effect of inducing greater similarity between polls where either the state or pollster is the same, but an alternative implementation gives qualitatively similar results.

3.6. More Uncertainties About the Similarity Between Elections

Overconfidence is a perennial problem for forecasting methods ([Lauderdale & Linzer, 2015](#)). As [Shirani-Mehr et al. \(2018\)](#) show using data similar to ours, the fundamental variance of poll estimates are larger than what an assumption of simple random sampling would suggest.

Here we relax our general premise that the patterns of survey error are similar between 2016 and 2020 by considering scenarios that reflect varying degrees of the uncertainty about the relevance of the historical lessons. Specifically, we extend our model to incorporate more uncertainty in our key assumption by introducing a (user-specified) inflation parameter λ to scale the variance of the γ_v . Intuitively, λ reflects how much we believe the similarity between 2016 and 2020 election polls (with respect to DDC* and WDC). That is, we generalize the first Normal model in (3.5) (which corresponds to setting $\lambda = 1$) to

$$(3.6) \quad \gamma_v \sim N(E_{mc}(\gamma_v^{pre}), \lambda \cdot Var_{mc}(\gamma_v^{pre})).$$

Although (3.6) does not explicitly account for improvement in 2020, by downweighting a highly unidirectional prior on the η , we achieve similar effect for our scenario analysis.

We remark in passing that our use of λ has the similar effect as adopting a power prior ([Ibrahim et al., 2015](#)), which uses a fractional exponent of the likelihood from *past data* as a prior for the current study. Both methods try to reduce the impact of the past data to reflect our uncertainties about their relevance to our current study. We use a series of λ 's to explore different degrees of this downweighting.

3.7. Sensitivity Check

There are many possible variations on the relatively simple framework presented here. In the Appendix we address two possible concerns with our approach: variation by pollster methodology and variation of the time window we average over. A natural concern is that our mix of pollsters might be masking vast heterogeneity in excess of that accounted for by performance in the 2016 election cycle. In Figure B2 we show separate results dividing pollsters into two groups, using FiveThirtyEight's pollster "grade" as a rough measure of pollster quality/methodology. We find that our model results do not vary by this distinction in most states, which is reasonable given that our overall model already accounts for house effects.

Separately, one might be concerned that averaging over a 3-month period might mask considerable change in actual opinion. In Figures B3 and B4, we therefore implement our method on polls conducted only in the last 3 weeks of our 2020 data collection period, a time window narrow enough to arguably capture a period where opinion does not change. Overall, we can draw qualitatively similar conclusions. Models from data in the last 3 weeks have similar point estimates and only slightly larger uncertainty estimates.

4. A Cautionary Tale

Our key result of interest is the posterior distribution of μ_i for each of the 12 battleground states. We compare our estimates with the baseline model that assumes no average bias, and inspect both the change in the point estimate and spread of the distribution.

Throughout, posterior draws are obtained using Hamiltonian Monte Carlo (HMC) as implemented in RStan ([Stan Development Team, 2020](#)). For each model, RStan produced 4 chains, each with 5,000 iterations, with the first 2,500 iterations discarded as burn-in. This resulted in 10,000 posterior draws for Monte Carlo estimations. Convergence was diagnosed by looking at traceplots and using the Gelman-Rubin diagnostic ([Gelman & Rubin, 1992](#)), with $R < 1.01$. For numerical stability, we multiply the DDC* in 2016 by 10^2 and WDC by 10^4 when performing all calculations. As noted in our modeling framework, the WDC approach has practical advantages over using DDC*, so we present only those results in the main text (analogous results for DDC* are shown in the Appendix).

Figure 1. Implications for modeling the 2016 Weighting Deficiency Coefficient (WDC) on 2020 state polls. Each histogram in each facet shows the posterior distribution of Trump's two-party voteshare in the state under our model. Blue distributions represent the baseline model (3.2) where we assume no selection bias in the reported survey topline, and the red ones are from the WDC model (3.3-3.5). The mean of each posterior distribution is given as text in each facet, in the same order. States are ordered by the mean of the baseline model.

Figure 1 shows for each state the posterior for Trump's voteshare. In blue, we show the posterior estimated from our baseline model assuming mean zero WDC (3.2). For ease of interpretation, states are ordered by the posterior mean of the baseline model, from most Democratic leaning to most Republican leaning. Note the tightness of the posterior distribution in the baseline model is a function of the total sample size: states that are polled more frequently have tighter distributions around the mean.

The estimates from our WDC models shift these distributions to the right, in most states. In Michigan, Wisconsin, and Pennsylvania, three 2016 Midwestern states Trump won, the posterior mean estimates moved about 1 percentage point from the baseline. In the median state, the WDC model (in red) moves the point estimate toward Trump by 0.8 percentage points. In North Carolina, Ohio, and Iowa this shift moves the posterior mean across or barely across the 50-50 line, which would determine the winner of the state. In states to the left of North Carolina according to the baseline model, the WDC models' point estimates indicate a likely Biden victory *even if* the polls suffer from selective responses comparable to those in 2016 polling. In Figure B2 in the Appendix, we also present the DDC* models and find that these are almost indistinguishable from the WDC models in most states. We caution that this is contingent on the assumption of turnout

remaining the same: for example, in Minnesota, if the 2020 total number of votes cast were 5 percent higher than in 2016, our WDC estimates would further shift away from the baseline model to further compensate for the 2016 WDC.

In general, as anticipated from our assumptions, the magnitude and direction of the shift roughly accords with the sample average errors as seen in 2016 (Table 2). Midwestern states that saw significant underestimates in 2016 exhibit an increase in Trump’s voteshare from our model, while states like Texas where polling errors were small tend to see a smaller shift from the baseline.

A key advantage of our results as opposed to simply adding back the polling error by a constant is that we capture uncertainty through a fully Bayesian model. The posterior distributions show that, in most states, the uncertainty of the estimate increases under our models, compared to under the baseline model. We further focus on this pattern in Figure 2, where we show how much the posterior standard deviation of a state’s estimate increases by our model over the baseline model, plotted against the state’s population for the purposes of our modeling. In the median state, the standard deviation of our WDC model output is about two times larger than that of the baseline. Figure 2 shows that this increase in uncertainty is systematically larger in higher populated states, such as Texas and Florida, than in lower populated states such as Iowa or New Hampshire. This is precisely what the Law of Large Population suggests [Meng \(2018\)](#), because our regressor X includes the population size N in the numerator.

It is important to stress that the uncertainties implied by our models only capture a particular sort of uncertainty: that based on the distribution of the DDC* or WDC that varies at the pollster level and state level. There are many other sources of uncertainties that can easily dominate, such as a change in public opinion among undecided voters late in the campaign cycle, a nonuniform shift in turnout, and a substantial change in the DDC*/WDC of pollsters from 2016.

While these uncertainties are essentially impossible to model, we can easily investigate the uncertainty in our belief of the degree of the recurrence of the 2016 surprise. Figure 3 shows the sensitivity of our scenario analysis results to the turning parameter λ introduced in Section 3.6 that indicates the relative weighting between the 2016 and 2020 data. For each value of $\lambda \in \{1, 1.5, \dots, 8\}$, we estimate the full posterior in each state as before and report three statistics: the proportion of draws in which Trump win’s two-party voteshare is over 50 percent (top panel), the posterior mean (bottom-left panel), and the posterior standard deviation (bottom-right panel).

Figure 2. The increase in uncertainty by state population (N). The vertical axis is a ratio of the standard deviation values taken from Figure 1, and serve as a measure of how much our model increases the uncertainty around our estimates.

We see that, as λ increases, so does the standard deviation. An eight-fold increase in λ leads to about a doubling in the spread of the posterior estimate. In most states (with the exception of Texas and Ohio), the increase in λ —which can be taken as a downweighting of the prior—also shifts the mean of the distribution against Trump by about 3 to 5 percentage points, and toward the baseline model. These trends consequently pull down Trump’s chances in likely Republican states, and bring the results of Texas and Ohio to tossups.

Figure 3. Sensitivity of WDC results to tuning parameter (λ). Estimates from our main model are rerun with various values of λ in (3.6). With higher values of λ , the estimated standard error of our posterior increases, and posterior means approach polls-only estimates assuming no bias. States are colored by their ratings by Cook Political Report as of September 17, 2020, where red indicates lean Republican, blue indicates lean Democratic, and green represents tossup.

5. Conclusion

Surveys are powerful tools to infer unknown population quantities from a relatively small sample of data, but the bias of an individual survey is usually unknown. In this article we show how, with hindsight, we can model the structure of this bias by incorporating fairly simple covariates such as population size. Methodologically, we demonstrate the advantage of using metrics such as DDC or WDC as a metric of bias in poll aggregation, rather than simply subtracting off point estimates of past error. In our application of the 2020 U.S. presidential election, our approach urges caution in interpreting a simple aggregation of polls. At the time of our writing (before Election Day), polls suggest a Biden lead in key battleground states. Our scenario analysis confirms

such leads in some states while casting doubts on others, providing a quantified reminder of a more nuanced outlook of the 2020 election.

Acknowledgments

We thank three anonymous reviewers for their thoughtful suggestions, which substantially improved the final article. We also thank Jonathan Robinson and Andrew Stone for their comments. We are indebted to Xiao-Li Meng for considerable advising on this project, detailed comments on improving the presentation of this article, and tokens of wisdom.

Disclosure Statement

HDSR Editor-in-Chief Xiao-Li Meng served an advisory role to the authors. The manuscript submission’s peer review process was blinded to him, and conducted by co-editors Liberty Vittert and Ryan Enos, who made all editorial decisions.

Appendix

Appendix A. Data Defect Correlation Model

A.1. Bayesian Beta Model for Modeling the Data Defect Correlation

Here, we describe how to modify the approach described in Section 3.2 to the case of Data Defect Correlation (DDC*, denoted by β_w). Since $|\rho| \leq 1$, we can map it into the interval $[0, 1]$, which then permits us to use the Beta distribution. We could use the usual Laplace or Normal distribution similar to our baseline model, but we prefer Beta to allow for the possibility of skewed and fatter-tailed distributions while still employing only two parameters. From the bounds established in [Meng.\(2018\)](#), we know the bounds on $|\rho|$ tend to be far more restrictive than 1; we will call this bound c . It is then easy to see that the transformation $B = 0.5(\rho + c)/c$ will map ρ into $(0, 1)$. Hence we can model B by a $\text{Beta}(\theta\phi, (1 - \theta)\phi)$ distribution, where θ is the mean parameter and ϕ is the shape parameter. This will then induce a distribution for ρ via $\rho = c(2B - 1)$. We pool the mean and shape across these groups on the logit and log scale, respectively, to transform our parameters to be unbounded.

Specifically, we assume:

$$\begin{aligned}
 \rho_{ijk}^{\text{pre}} | \theta_{ij}, \phi_{ij} &\sim c [2 \cdot \text{Beta}(\theta_{ij}\phi_{ij}, (1 - \theta_{ij})\phi_{ij}) - 1], \\
 (A1) \quad \theta_{ij} &= \text{logit}^{-1} \left(\gamma_0 + \gamma_i^{\text{state}} + \gamma_j^{\text{pollster}} \right), \\
 \phi_{ij} &= \exp \left(\psi_0 + \psi_i^{\text{state}} + \psi_j^{\text{pollster}} \right),
 \end{aligned}$$

where γ, ψ are given hierarchical normal priors with mean 0 and variance τ_*^2 in the same manner as the weighting deficiency coefficient (WDC). The prior on τ_*^2 is the same as in the main text as well, that is, $\text{gamma}(5, 20)$. We set $c = 0.011$ because over the course of the past five presidential elections, ρ_{ijk} at the state-level varies well within a range of about -0.01 to 0.01 . So, our choice of c is a safe bound.

Under this alternative parameterization of the Beta distribution, the θ_{ij} can be interpreted as the mean DDC for pollster j conducting surveys in state i . The ϕ_{ij} is a shape parameter that determines the variations in DDC in that state-pollster pair, much like σ_{ij} in the original formulation.

A.2. Simulating 2020

In order to simulate the 2020 elections, we mimic our approach in (3.4) to model:

$$\begin{aligned}
 \sqrt{X_{ijk}} [\hat{\delta}_{ijk} - \delta_i] | \theta_{ij}, \phi_{ij} &\sim c [2 \cdot \text{Beta}(\theta_{ij} \phi_{ij}, (1 - \theta_{ij}) \phi_{ij}) - 1], \\
 \theta_{ij} &= \text{logit}^{-1} \left(\gamma_0 + \gamma_i^{\text{state}} + \gamma_j^{\text{pollster}} \right), \\
 \phi_{ij} &= \exp \left(\psi_0 + \psi_i^{\text{state}} + \psi_j^{\text{pollster}} \right), \\
 \gamma_v &\sim N(E_{\text{mc}}(\gamma_v^{\text{pre}}), \text{Var}_{\text{mc}}(\gamma_v^{\text{pre}})), \\
 \psi_v &\sim N(E_{\text{mc}}(\psi_v^{\text{pre}}), \text{Var}_{\text{mc}}(\psi_v^{\text{pre}})), \\
 \delta_i | \mu_\delta, \tau_\delta &\stackrel{\text{i.i.d.}}{\sim} N(\mu_\delta, \tau_\delta^2).
 \end{aligned}
 \tag{A2}$$

The results on the sensitivity to λ with this Beta model are shown in Figure A1.

Figure A1. Sensitivity of DDC* model to tuning parameter. Same as Figure 3 but using DDC*.

Appendix B: Additional Results

Table B1. Pollster grades.

	538 Grade	Polls
Ipsos	B-	24
YouGov	B	23
New York Times / Siena	A+	18

Emerson College	A-	13
Quinnipiac University	B+	12
Monmouth University	A+	11
Public Policy Polling	B	11
Rasmussen Reports	C+	8
Suffolk University	A	6
SurveyUSA	A	6
CNN / SSRS	B/C	4
Marist College	A+	4
Data Orbital	A/B	3
Marquette University	A/B	3
University of New Hampshire	B-	2
Gravis Marketing	C	1
Mitchell Research & Communications	C-	1

B.1. DDC* and WDC Models

Figure B2 shows the DDC models (in orange) in addition to the WDC models (in red). In the center and right columns, we also show results separating pollsters by FiveThirtyEight’s pollster “grade.” The grade is computed by FiveThirtyEight incorporating the historical polling error of the pollsters, and is listed in Table B1. When modeling 2016 error, we retroactively assign these grades to the pollster instead of using the pollster’s 2016 grade.

B.2. Only Using Data 3 Weeks Out

To address the concern that using data from 3 months prior to the election would smooth out actual changes in the underlying voter support, we redid our analysis with the last 3 weeks of our data, from October 1 to 21, 2020. This resulted in less than half the number of polls (72 polls in the same 12 states, but covered by 15 pollsters).

Figure B3 shows the same posterior estimates as Figure 1 but using this subset. We see that the model estimates are largely similar to those from data using 3 months. To more clearly show the key differences, we compare the summary statistics of the two set of estimates directly in the scatterplot in Figure B4. This confirms that the posterior means of the estimates do not change even after halving the size of the data set. The standard deviation of the estimates increases, as expected, but only slightly.

Figure B2. DDC* and WDC models. In the first column, we show estimates from all pollsters, as in the main text, showing results for both the DDC* and WDC models. In the next two columns, we do the same but subsetting pollsters in two, by their FiveThirtyEight grade.

Figure B3. Replicating Figure 1 with 3 weeks worth of data.

Each point represents a summary statistic (a mean or a standard deviation) state-model combination.

Figure B4. Comparison of summary statistics with 3 week subset.

Data Repository / Code

The authors' code and data are available for public access on CodeOcean, here:

<https://doi.org/10.24433/CO.6312350.v1>

References

- Caughey, D., Berinsky, A. J., Chatfield, S., Hartman, E., Schickler, E., & Sekhon, J. S. (2020). *Target estimation and adjustment weighting for survey nonresponse and sampling bias*. Cambridge University Press. <https://doi.org/10.1017/9781108879217>
- Cohn, N. (2018). Two vastly different election outcomes that hinge on a few dozen close contests. *New York Times*. <https://perma.cc/NYR4-H45G>
- Cohn, N. (2018). What the polls got right this year, and where they went wrong. *New York Times*. <https://perma.cc/HFT3-XW8W>
- 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(4), 409–451. <https://doi.org/10.1017/S0007123400006682>
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. <https://doi.org/10.1214/ss/1177011136>
- Gelman, A., Goel, S., Rivers, D., & Rothschild, D. (2016). The mythical swing voter. *Quarterly Journal of Political Science*, 11(1), 103–130. <https://doi.org/10.1561/100.0001503>
- Hummel, P., & Rothschild, D. (2014). Fundamental models for forecasting elections at the state level. *Electoral Studies*, 35, 123–139. <https://doi.org/10.1016/j.electstud.2014.05.002>
- Ibrahim, J. G., Chen, M.-H., Gwon, Y., & Chen, F. (2015). The power prior: Theory and applications. *Statistics in Medicine*, 34(28), 3724–3749. <https://doi.org/10.1002/sim.6728>
- Jackman, S. (2014). The predictive power of uniform swing. *PS: Political Science & Politics*, 47(2), 317–321. <https://doi.org/10.1017/S1049096514000109>
- Jennings, W., & Wlezien, C. (2018). Election polling errors across time and space. *Nature Human Behaviour*, 2(4), 276–283. <https://doi.org/10.1038/s41562-018-0315-6>
- Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., & Rivers, D. (2018). An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, 82(1), 1–33. <https://doi.org/10.1093/poq/nfx047>

Kish, L. (1965). *Survey sampling*. John Wiley and Sons.

Kuriwaki, S. (2020). *ddi: The data defect index for samples that may not be IID*. <https://CRAN.R-project.org/package=ddi>

Lauderdale, B. E., & Linzer, D. (2015). Under-performing, over-performing, or just performing? The limitations of fundamentals-based presidential election forecasting. *International Journal of Forecasting*, 31(3), 965–979. <https://doi.org/10.1016/j.ijforecast.2015.03.002>

Linzer, D. A. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, 108(501), 124–134. <https://doi.org/10.1080/01621459.2012.737735>

McCarthy, Justin. (2019). *High enthusiasm about voting in U.S. heading into 2020*. <https://perma.cc/UFK6-LWR9>

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2), 685–726. <https://doi.org/10.1214/18-AOAS1161SF>

Rentsch, A., Schaffner, B. F., & Gross, J. H. (2019). The elusive likely voter: Improving electoral predictions with more informed vote-propensity models. *Public Opinion Quarterly*, 83(4), 782–804. <https://doi.org/10.1093/poq/nfz052>

Shirani-Mehr, H., Rothschild, D., Goel, S., & Gelman, A. (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association*, 113(522), 607–614. <https://doi.org/10.1080/01621459.2018.1448823>

Skelley, G., & Rakich, N. (2020). What pollsters have changed since 2016 — And what still worries them about 2020. *FiveThirtyEight.Com*. <https://perma.cc/E4FC-6WSL>

Stan Development Team\phantom. (2020). *Rstan: The R Interface to Stan*. <http://mc-stan.org>

Wright, F. A., & Wright, A. A. (2018). How surprising was Trump’s victory? Evaluations of the 2016 US presidential election and a new poll aggregation model. *Electoral Studies*, 54, 81–89. <https://doi.org/10.1016/j.electstud.2018.05.001>

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75(4), 709–747. <https://doi.org/10.1093/poq/nfr020>

Yu, K., & Moyeed, R. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4), 437–447. [https://doi.org/10.1016/S0167-7152\(01\)00124-9](https://doi.org/10.1016/S0167-7152(01)00124-9)

©2020 Michael Isakov and Shiro Kuriwaki. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.

References

- Stan Development Team. (2020). *Rstan: The R Interface to Stan*. <http://mc-stan.org>
- Caughey, D., Berinsky, A. J., Chatfield, S., Hartman, E., Schickler, E., & Sekhon, J. S. (2020). *Target Estimation and Adjustment Weighting for Survey Nonresponse and Sampling Bias*. Cambridge University Press. <https://doi.org/10.1017/9781108879217>
- Cohn, N. (2018). Two Vastly Different Election Outcomes that Hinge on a Few Dozen Close Contests. *New York Times*. <https://perma.cc/NYR4-H45G>
- Cohn, N. (2018). What the Polls Got Right This Year, and Where They Went Wrong. *New York Times*. <https://perma.cc/HFT3-XW8W>
- 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(4), 409–451. <https://doi.org/10.1017/S0007123400006682>
- Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4), 457–472. <https://doi.org/10.1214/ss/1177011136>
- Gelman, A., Goel, S., Rivers, D., & Rothschild, D. (2016). The Mythical Swing Voter. *Quarterly Journal of Political Science*, 11(1), 103–130. <https://doi.org/10.1561/100.0001503>
- Hummel, P., & Rothschild, D. (2014). Fundamental Models for Forecasting Elections at the State Level. *Electoral Studies*, 35, 123–139. <https://doi.org/10.1016/j.electstud.2014.05.002>
- Ibrahim, J. G., Chen, M.-H., Gwon, Y., & Chen, F. (2015). The Power Prior: Theory and Applications. *Statistics in Medicine*, 34(28), 3724–3749. <https://doi.org/10.1002/sim.6728>
- Jackman, S. (2014). The Predictive Power of Uniform Swing. *PS: Political Science & Politics*, 47(2), 317–321. <https://doi.org/10.1017/S1049096514000109>
- Jennings, W., & Wlezien, C. (2018). Election Polling Errors across Time and Space. *Nature Human Behaviour*, 2(4), 276–283. <https://doi.org/10.1038/s41562-018-0315-6>
- Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., & Rivers, D. (2018). An Evaluation of the 2016 Election Polls in the United States. *Public Opinion Quarterly*, 82(1), 1–33. <https://doi.org/10.1093/poq/nfx047>
- Kish, L. (1965). *Survey Sampling*. John Wiley.

- Kuriwaki, S. (2020). *ddi: The Data Defect Index for Samples that May not be IID*. <https://CRAN.R-project.org/package=ddi> ↵
- Lauderdale, B. E., & Linzer, D. (2015). Under-performing, Over-performing, or Just Performing? The Limitations of fundamentals-based Presidential Election Forecasting. *International Journal of Forecasting*, 31(3), 965–979. <https://doi.org/10.1016/j.ijforecast.2015.03.002> ↵
- Linzer, D. A. (2013). Dynamic Bayesian Forecasting of Presidential Elections in the States. *Journal of the American Statistical Association*, 108(501), 124–134. <https://doi.org/10.1080/01621459.2012.737735> ↵
- McCarthy, Justin. (2019). *High Enthusiasm About Voting in U.S. Heading Into 2020*. <https://perma.cc/UFK6-LWR9> ↵
- Meng, X.-L. (2018). Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election. *The Annals of Applied Statistics*, 12(2), 685–726. <https://doi.org/10.1214/18-AOAS1161SF> ↵
- Rentsch, A., Schaffner, B. F., & Gross, J. H. (2019). The Elusive Likely Voter: Improving Electoral Predictions With More Informed Vote-Propensity Models. *Public Opinion Quarterly*, 83(4), 782–804. <https://doi.org/10.1093/poq/nfz052> ↵
- Shirani-Mehr, H., Rothschild, D., Goel, S., & Gelman, A. (2018). Disentangling Bias and Variance in Election Polls. *Journal of the American Statistical Association*, 113(522), 607–614. <https://doi.org/10.1080/01621459.2018.1448823> ↵
- Skelley, G., & Rakich, N. (2020). What Pollsters Have Changed Since 2016 — And What Still Worries Them About 2020. *FiveThirtyEight.com*. <https://perma.cc/E4FC-6WSL> ↵
- Wright, F. A., & Wright, A. A. (2018). How Surprising Was Trump’s Victory? Evaluations of the 2016 US Presidential Election and a New Poll Aggregation Model. *Electoral Studies*, 54, 81–89. <https://doi.org/10.1016/j.electstud.2018.05.001> ↵
- Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-probability Samples. *Public Opinion Quarterly*, 75(4), 709–747. <https://doi.org/10.1093/poq/nfr020> ↵
- Yu, K., & Moyeed, R. (2001). Bayesian Quantile Regression. *Statistics & Probability Letters*, 54(4), 437–447. [https://doi.org/10.1016/S0167-7152\(01\)00124-9](https://doi.org/10.1016/S0167-7152(01)00124-9) ↵