SUPPLEMENTARY ONLINE MATERIALS FOR:

AN EVALUATION OF THE 2016 ELECTION POLLS IN THE UNITED STATES

Courtney Kennedy (corresponding author), Pew Research Center, ckennedy@pewresearch.org

Mark Blumenthal, SurveyMonkey

Scott Clement, Washington Post

Joshua D. Clinton, Vanderbilt University

Claire Durand, University of Montreal

Charles Franklin, Marquette University

Kyley McGeeney, PSB

Lee Miringoff, Marist College

Kristen Olson, University of Nebraska-Lincoln

Doug Rivers, Stanford University, YouGov

Lydia Saad, Gallup

Evans Witt, Princeton Survey Research Associates

Christopher Wlezien, University of Texas at Austin

SECTIONS

Appendix A: List of Microdatasets Made Available to the Committee

This appendix presents information about poll microdatasets that were made available to the committee for its analysis. Most of the datasets are available from the Roper Center or individual polling organizations' websites (e.g., USC Dornsife/LA Times and Pew Research Center).

Table A.1 Microdatasets Made Available to the Committee

Organization	Microdata
ABC News/ Washington Post	National tracking poll with n=9,930 fielded Oct 20-Nov 7
CNN/ORC	National poll with n=1,001 fielded Sep 1-4; National poll with n=1,501 fielded Sep 28-Oct 2; National poll with n=1,017 fielded Oct 20-23; AZ poll with n=1,005 fielded Oct 27-Nov1; CO poll with n=1,009 fielded Sep 20-25; FL poll with n=1,000 fielded Sep 7-12; FL poll with n=1,011 fielded Oct 27-Nov 1; NC poll with n=1,025 fielded Oct 10-15; NV poll with n=1,006 fielded Oct 10-15; NV poll with n=1,005 fielded Oct 27-Nov 1; OH poll with n=1,004 fielded Sep 7-12; OH poll with n=1,008 fielded Oct 10-15; PA poll with n=1,032 fielded Sep 20-25; PA poll with n=1,014 fielded Oct 27-Nov 1
Marquette University	WI state poll with 1,401 fielded Oct 27-30
Michigan State University	MI state poll with n=1,010 fielded Sep 1-Nov 13
Monmouth University	NV state poll with n=465 fielded Oct 14-17; WI state poll with n=428 fielded Oct 15-18; NC state poll with n=487 fielded Oct 20-23; AZ state poll with n=463 fielded Oct 21-24; NH state poll with n=430 fielded Oct 22-25; IN state poll with n=448 fielded Oct 27-30; MO state poll with n=457 fielded Oct 28-31; PA state poll with n=453 fielded Oct 29-Nov 1; UT state poll with n=445 fielded Oct 30-Nov 2
Pew Research Center	Election Callback Study 2000, 2004, 2008, 2012, 2016; Cumulative national polls from 2016 with cumulative n=15,812; 2016 Callback Study n=1,254 fielded Nov 10-14, 2016
SurveyMonkey	National tracking poll with n=219,431 fielded Oct 4-Nov 7. This dataset also supported state-level analyses.
USC Dornsife/LA Times	2016 National panel survey, 4,509 fielded Jul 4-Nov 7
YouGov	Cooperative Congressional Election Study with n=117,123 fielded Oct 4-Nov 6; Economist/YouGov poll with n=4,171 fielded Nov 4-7; Other polls across 51 states with n=81,246 fielded Oct 24-Nov 6. These datasets would have supported state-level analyses. No weights were provided.

Appendix B: Additional Analysis of Polling Errors in the 2016 General Election

This appendix contains additional details of the committee's analysis of the accuracy of polling errors in the 2016 general election. Table B.1 reports the average absolute and signed errors in 2016 general election polls conducted nationally, in battleground states, and non-battleground states.

Table B.1 Average Absolute and Average Signed Error in 2016 State-Level General Election Polls

Type of poll	Number of polls in final 13 days	Average absolute error	Average signed error
National polls	39	2.1	1.3
All state polls	422	5.1	3.0
All battleground state polls	206	3.6	2.3
All non-battleground polls	216	6.5	3.6
Wisconsin	13	6.5	6.5
Ohio	14	4.8	4.8
Minnesota	5	4.9	4.9
Pennsylvania	24	4.2	4.2
North Carolina	18	4.8	4.0
Michigan	16	3.8	3.5
New Hampshire	16	5.0	3.4
Florida	23	2.9	1.3
Arizona	17	2.6	1.1
Georgia	14	2.3	0.9
Virginia	15	2.0	-0.3
Colorado	16	2.3	-1.6
Nevada	15	2.5	-1.7

Table B.2 presents the OLS model regressing absolute error in general election polls on poll characteristics. The results corroborate the bivariate finding that polls using IVR tended to have less error in the 2016 general election (Figure 4 in the text).

Table B.2 Regression of Absolute Error on Poll Characteristics

	Model 1 (Mode)			Model 2 (Sample)		
	B	Sig	S.E.	*B*	Sig	S.E.
(Intercept)	1.52	**	0.535	1.62785	**	0.598
Mode						
Internet	0.19		0.401			
IVR	-1.14	*	0.553			
IVR/Cell	-0.38		0.628			
IVR/Internet	-0.39		0.486			
Other	4.78	***	1.244			
Sample source						
Opt-in				-0.12		0.494
Voter file				-0.56		0.552
Opt-in/Voter file				-0.29		0.635
RDD/Opt-in				-0.93		1.168
Voter file/RDD				1.17		1.765
Other				-0.82		0.865
Arizona	0.76		0.657	0.86		0.719
Colorado	0.70		0.689	0.69		0.765
Florida	1.18	.	0.618	1.29	.	0.672
Georgia	0.79		0.725	0.73		0.771
Michigan	2.49	***	0.714	2.20	**	0.763
Minnesota	2.93	**	1.104	3.08	**	1.164
Missouri	2.75		2.33	1.41		2.965
North Carolina	3.24	***	0.665	3.17	***	0.714
New Hampshire	3.38	***	0.687	3.27	***	0.735
Nevada	0.91		0.707	0.87		0.764
Ohio	2.52	***	0.749	3.17	***	0.782
Pennsylvania	2.43	***	0.613	2.51	***	0.667
Virginia	0.27		0.72	0.27		0.78
Wisconsin	4.87	***	0.749	4.85	***	0.807
Days from mid-date to election	0.03		0.046	0.04		0.049
Adjusted R-Squared	.28			.27		

Reference categories: Live phone (Mode), RDD (Sample source), National (geography).

Figure B.1 shows average absolute error in national general election polls, by design, in recent elections. The results indicate that IVR-only polls tended to fare worse than other modes in both 2008 and 2012.

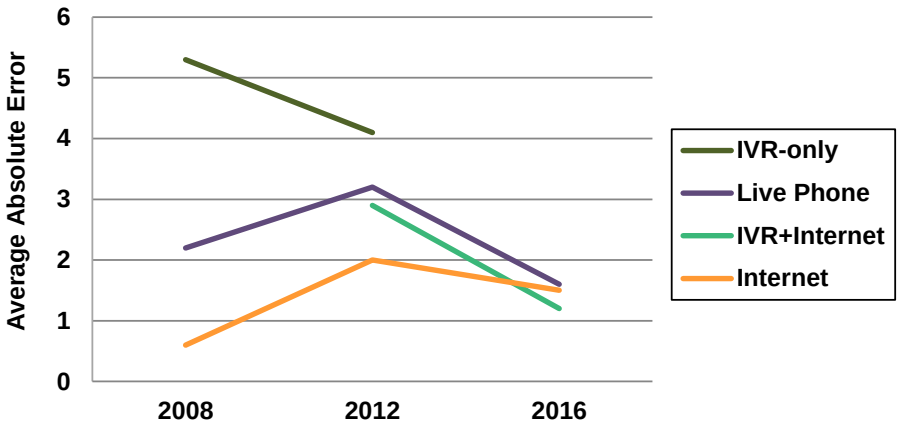

Figure B.1. Absolute Average Error in National Polls by Mode by Year.
Note – In 2016 there were no national polls conducted using only IVR.

Appendix C: Additional Analysis of Polling Errors in the 2016 Primaries and Caucuses

This appendix contains additional details of the committee's analysis of the accuracy of polling errors in the 2016 primaries and caucuses.

C.1 Regression Analysis Examining Effects of Poll Design Features on Accuracy

The focus of this section is on the average overall performance of the polls in a state primary or caucus – not the performance of individual polls or even specific types of polls. Our motivating question is – among the polls conducted and publicly reported in the last two weeks for each contest, how well did the polls do at predicting the margin of victory in each contest on average? Are there characteristics of polls or contests that are related to better or worse performance on average? To do so, we collect information on all publicly reported polls conducted within the last two weeks of each primary contest and reported by FiveThirtyEight.com, Pollster.com.

To explore differences in polling performance, for each poll we collected the following: the length of the field period, the firm conducting the poll, the sample size, the target population ("likely" voter or registered voter), the interview mode, the sample source (when possible), the percentage of cell phones in the sample (when possible), the affiliation of the pollster (partisan, sponsored, or nonpartisan), the votes received by each of the leading candidates, and the verified election results for each contest.

There was very little variation for some of these characteristics. Because 441 of the surveys had a target population of "likely voters" and only 14 reported results of registered voters in the time frame we examine, for example, we have no real ability to determine whether likely voter or registered voter samples are more accurate. Other data was hard to collect – even

after trying to contact every pollster we were only able to acquire the percentage of cell phone numbers called for 323 of the publicly available polls.

While the performance of 2016 primary polls seems relatively consistent with the performance of polls in earlier primary contests, to delve deeper into the data and to characterize how polling performance varies across the primary contests in 2016, we examine how the median absolute polling error varies by the number of polls being conducted in a state's Democratic and Republican primaries. We focus on the median absolute polling error to minimize the impact of extreme outliers, but the takeaways are unchanged. Figure C.1 presents the performance of polls within each primary contest.

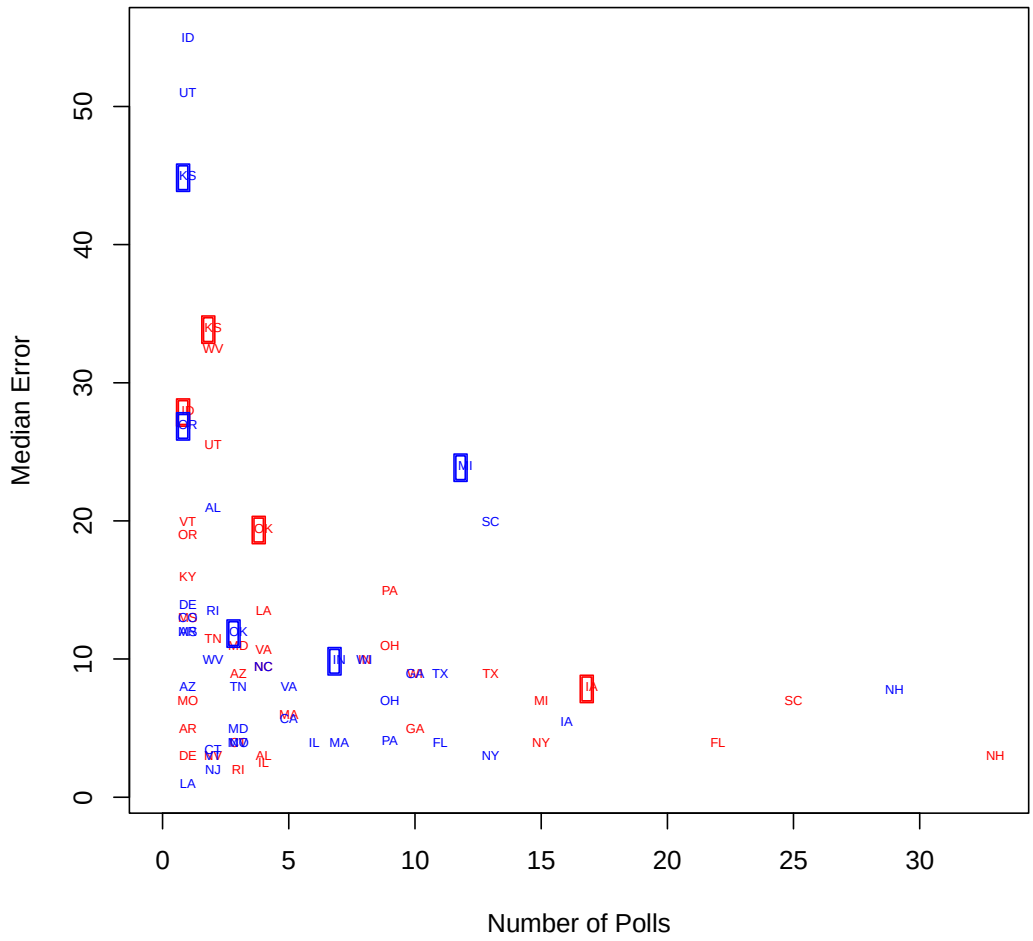

Figure C.1 Median Absolute Error in Primary Contests: Note – Republican (red) and Democrat (blue) results for each state are plotted.

Each labeled point in Figure C.1 denotes the median absolute error for the polls conducted in each state contest for the Republicans (red) and Democrats (blue) as a function of the number of polls. Circled states indicate instances in which more than 50% of the polls predicted the wrong winner – something that happened in 9 out of the 78 contests.

Several conclusions are evident from Figure C.1. First, the number of polls conducted in contests varies considerably – ranging from a high in the New Hampshire Republican primary of 33 polls to a low of a single poll in 19 contests. This variation is important for several reasons.

First, insofar as each poll result is an independent estimate of the result, the average absolute error in a contest should be smaller the more polls there are for the same reason that more respondents in a poll lead to a smaller margin of error all else equal. Of course, the polls being averaged vary in important ways that can undermine the assumption that the polls' estimates are a random sample of the population, but the fact that a smaller average error occurs in states with more polls suggests that there are more similarities than differences. (Note that this relationship is not necessarily evidence of "herding," whereby polls are weighted to help mimic pre-existing results; if herding occurs, there is no reason to think that it would necessarily be more prevalent in states with more polls.)

Second, the variation we observe in the number of polls in each contest highlights an important limitation to our efforts to evaluate the accuracy of polls. Because each pollster decides which contests to poll, this choice can have important implications for evaluating the overall accuracy of polls. If the decision of whether or not to poll depends on the difficulty of polling in the state, the fact that only some pollsters choose to poll a contest can affect our overall assessment of poll quality. To use an analogy, evaluating the accuracy of polls using their performance in the states pollsters choose to poll in is akin to evaluating a student's

performance on a test using only those questions that they choose to answer. If students decide to only answer "easy" questions, our evaluation of their ability may be very misleading. Similarly, if pollsters are more likely to poll in states that they are more likely to be successful in, our assessment may be overly optimistic. As a result, our results can, at best, inform us of how well the polls that were conducted and publicly released performed in those states where they were conducted. Because not every pollster polls every race and the decision to poll or not to poll – or to perhaps to publicly release the poll results or not – our results could be affected by the difficulty of polling the race itself if polls are more likely to be conducted in easier states to poll in.

Finally, highlighting a point made earlier, the median of the median absolute error across the 78 contests with at least one poll conducted in the last two weeks is 9.0. That is the median amount of error between the estimated and actual margin of victory across all primary contests is 9 points. Thus, while the polls correctly predicted the winner more often than not, on average, the predicted margin of victory in polls was nine points different than the official margin on Election Day.

To analyze poll performance based on their characteristics, we estimate the absolute value of each poll's error as a function of both poll-level and contest-level characteristics using a linear regression model. The benefit of this approach is that it allows us to directly quantify the average conditional impact of each characteristic holding all other aspects of the poll and contest fixed. This approach provides a high-level overview of the features that are related to larger and smaller errors while quantifying the average overall performance.

To do so, we control for several contest level features, including: whether it is a Republican or Democratic contest (perhaps it is harder to predict the margin when more candidates are running?), the state in which the contest occurs (to control for potential differences in the difficulty of polling in different states), the total number of polls that were conducted in the primary contest in the state (to provide a sense of how much other activity was going on in the contest), and the percentage of the vote received by the winner (perhaps it is harder to predict the margin in blow-out contests that in closely fought contests?).

We also account for several poll-specific characteristics that may affect the accuracy of the poll. The variables we control for include: the sample size (and the square of the sample size to allow for a non-linear effect), the length of the field period (and its' square) to account for the potential impact of larger and smaller field periods, the number of days between the last field period day and Election Day to account for the possibility that later polls may be more accurate because they capture last minute changes in opinion, and whether the pollster is affiliated with the Democratic or Republican party. To allow for potential expertise effects, we also interact the partisanship of pollsters with the party of the contest to explore whether Democratic Pollsters are more accurate in Democratic primaries, for example.

The final set of variables involve the mode of survey interview and whether it was done via: interactive voice response (IVR) (86), IVR/Live Phone (9), IVR/Live Phone/Online (3), IVR/Online (66), Internet (47), Live Phone (239) or Live Phone/Online (6). We collapse these into a set of three non-exclusive, but exhaustive variables depending on whether the poll relies either exclusively or partially on each of the three modes. Given the interest in differences by polling mode, Figure C.2 presents the distribution of polling errors by mode.

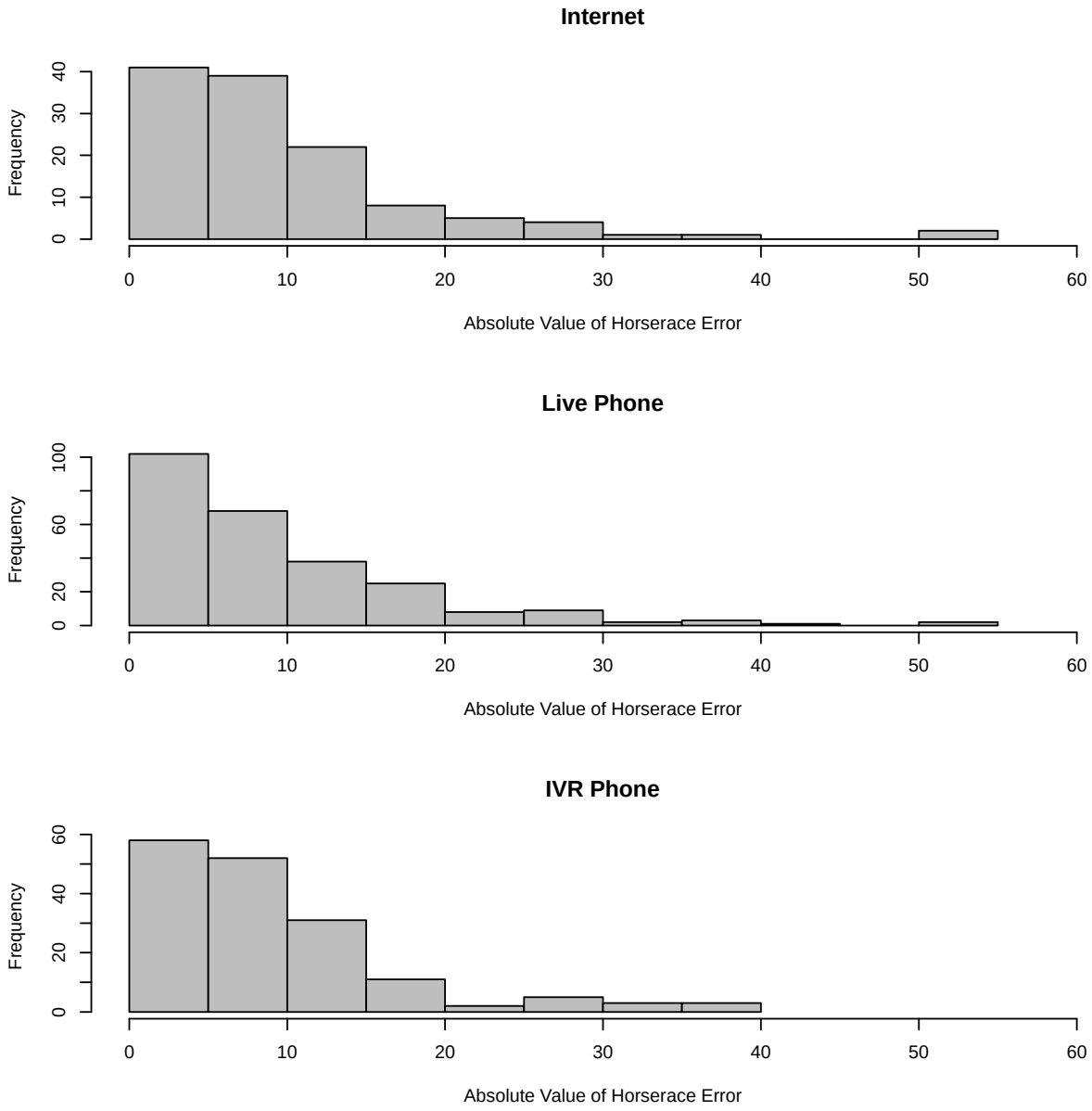

Figure C.2 Absolute Error by Mode

Figure C.2 reveals that there are few differences in the absolute error when we look at the impact of survey mode – the median horserace error for internet polls, live phone polls and IVR polls are 8%, 7% and 8%. Even so, it is hard to make direct comparisons because not only are there differences in how polls are being conducted within each mode, but also not every mode is being used for every primary. Some primaries – typically primaries for which one candidate was heavily favored – lacked a single live phone poll, and if the margin of victory in these primaries

are harder to predict this would impact our ability to interpret these differences as reflecting the impact of survey mode.

To better explain the relative performance of polls it is, therefore, important to control for as many aspects as possible to allow us to make a comparison, "all else equal." We use a regression specification that includes the characteristics described above to do so. The results of this are perhaps best digested graphically. Figure C.3 depicts the coefficient estimate and the 95% confidence intervals for the survey characteristics we are able to include in the analysis, given data constraints. Several conclusions are immediately evident. First, while there are slight differences by survey mode – polls using IVR and online methods are associated with slightly larger average absolute errors, all else equal, the differences are small (.21 and .08 larger than a phone poll, respectively) the differences are not statistically distinguishable from 0. However, polls conducted further from the election contain a larger error – for every day difference between Election Day and the last field period, the average error is 0.40 larger, all else equal.

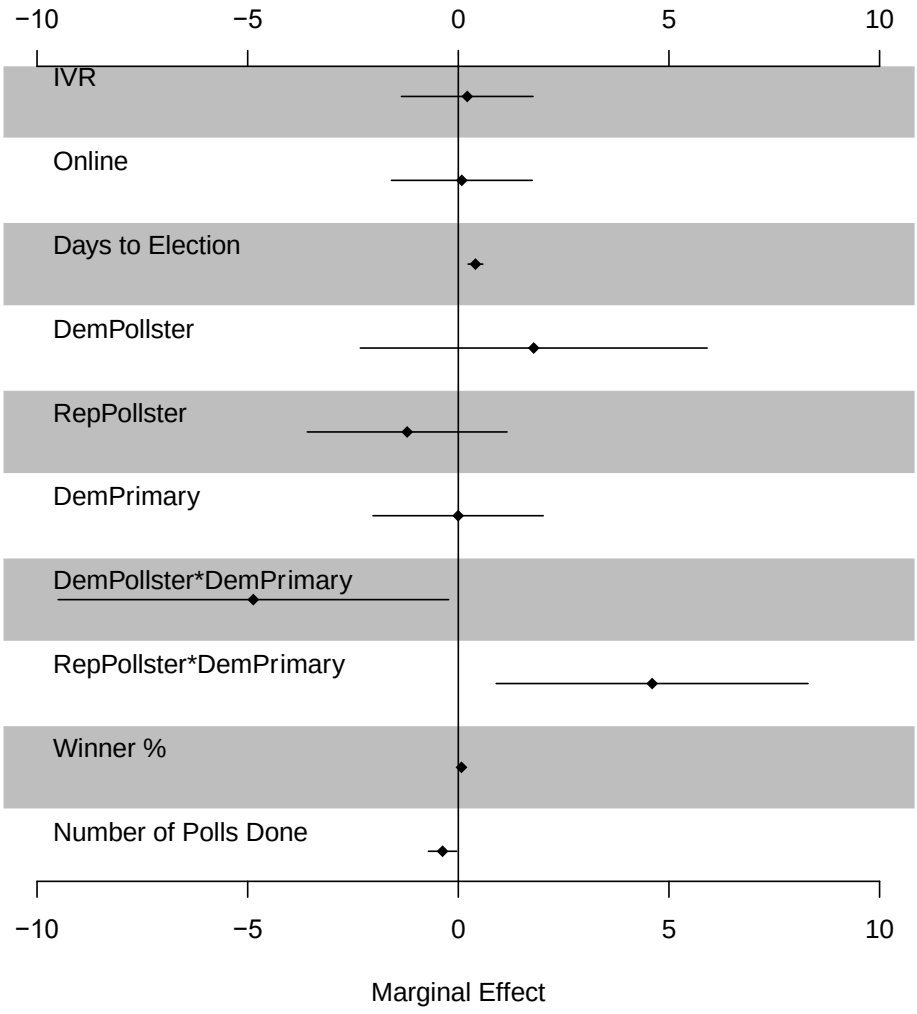

Figure C.3 Marginal Effect of a One-Unit Change in Each Feature on the Absolute Error for 2016 Primary Polls

The partisanship of the pollster also seems to have an interesting effect. Because nonpartisan pollsters are the omitted category, the impact of *DemPollster* and *RepPollster* reflects the relative performance of Democratic pollsters and Republican pollsters, respectively, in a Republican primary contest compared to a nonpartisan poll. While not distinguishable from 0 at conventional levels, the estimates suggest that Democratic pollsters' error is 1.79 larger than nonpartisan pollsters while Republican pollsters are 1.22 smaller. The opposite pattern emerges when we look at the performance of partisan pollsters in a Democratic contest. In such cases,

Democratic pollsters make errors that are 4.87 smaller on average than a nonpartisan pollster and Republican pollsters make errors that are 4.60 larger. The fact that the performance of partisan pollsters varies, and it is smaller in the primary contests that match the pollsters' affiliation suggests that perhaps partisan pollsters may have a slightly better ability to predict their own contests – a disparity that is most striking in Democratic contests.[1] That said, it is important to emphasize that this difference is driven by the performance of a few pollsters in a few contests, so it is important to not over-interpret the significance of this finding.

There is also important variation in average poll performance depending on whether the election is a blowout or not, as well as the number of polls that are conducted in the state. While distinguishable from zero, the substantive magnitude of the electoral margin on poll performance is relatively slight – increasing the margin of victory by a standard deviation (12.4 points) is predicted to increase the average horserace error by 0.84 all else equal. Similarly, while the average polling error is smaller in contests with more polls, the effect size of -0.38 suggests a substantively slight impact – going from a contest with a single poll to a contest with 33 polls conducted in the last two weeks is associated with a decrease of only 1.27 in the average absolute horserace margin of error.

Of course, there are also systemic effects that may vary by state. Not every state is equally easy to poll in, and in estimating the effect for each characteristic we also control for differences across states. These differences sometimes matter. Polls in Utah, South Carolina, Oregon, Michigan and Kansas, for example, were all off by an average of 10% all else equal.

[1] Note that there are 36 polls by a Democratic pollster in the sample and 24 are taken in a Democratic primary contest. These 24 polls were all done by PPP using an IVR methodology. There are 60 polls by a Republican affiliated pollster, and 41 of those are taken in a Republican primary. Republican polls in Democratic contests were done by a variety of firms including: Gravis, Magellan Strategies, TargetPoint, Landmark and Mitchell.

While it is impossible for us to diagnose the exact reasons for these systematic errors, controlling for them in the analysis is important because it removes the impact of these state-specific errors from the estimated effects graphed in Figure C.1.

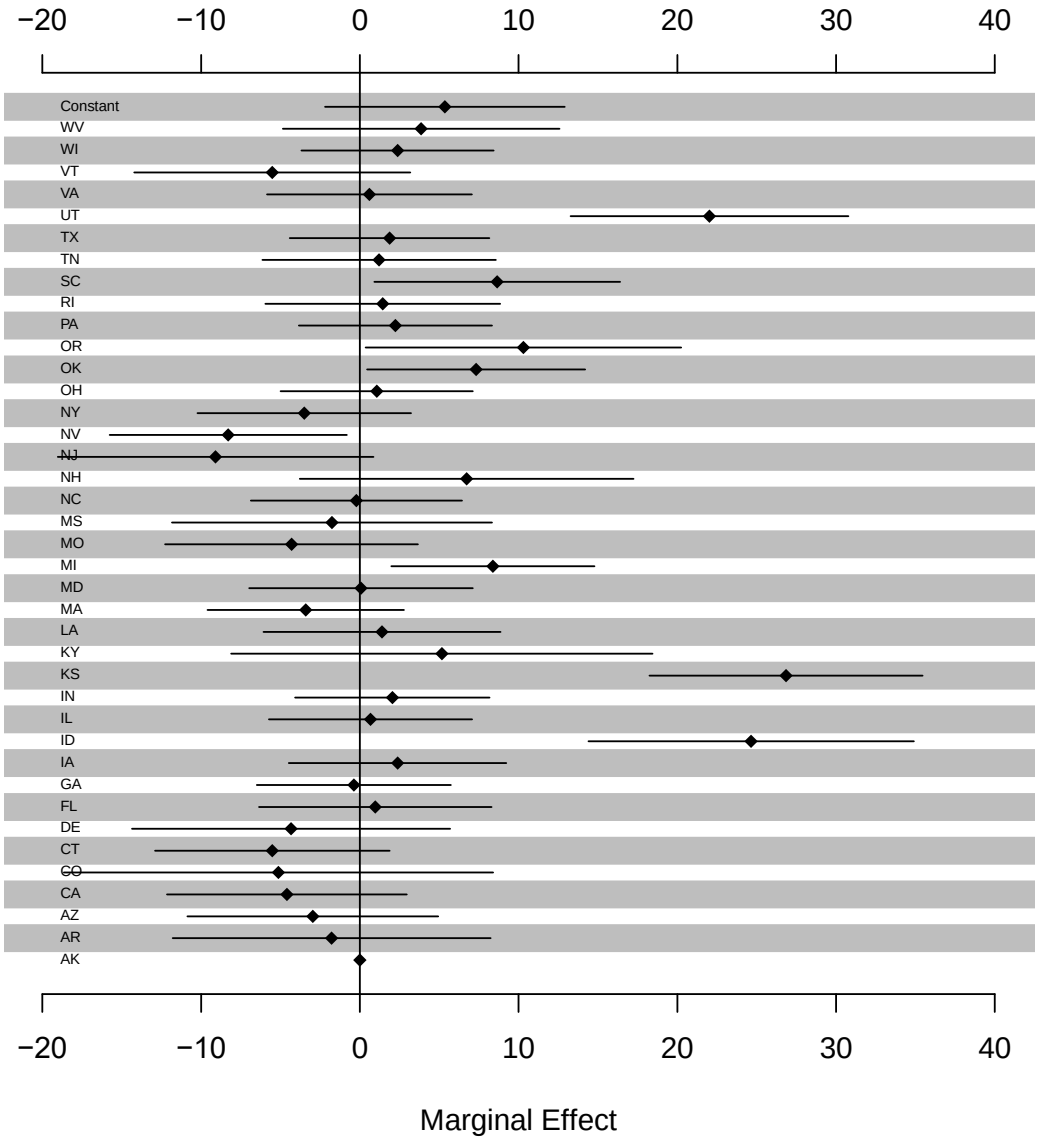

Figure C.4 Average Error in Polls After Controlling for Poll Characteristics

The value of the constant is substantively important, as it reflects the average amount of error in the polls' horserace estimates after controlling for poll-level and contest-level differences. The

estimate is 5.3 with a 95% confidence interval ranging from -2.2 to 12.9. This means that while the average estimate of the margin of victory was off by nearly 5 points, we cannot statistically reject the hypothesis that the average error was 0 at conventional significance levels.

What explains the variation in performance across states? To tackle this question we see what characteristics predict the average absolute horserace error in each of the 74 state primary contests in which at least one poll was taken in the two weeks prior to the election. We collect data on whether the primary contest is closed, open, or mixed, whether it is a caucus, whether it is a Republican or Democratic primary, how many votes were cast in the election (logged to account for outliers) and the number of polls that were conducted.

Table C.1 Results of Regressing Absolute Poll Error on Contest Characteristics

	Coefficient (Standard Error)
Closed Primary	-1.64
	(3.41)
Open Primary	2.51
	(3.31)
Caucus	9.68*
	(3.71)
Republican Contest	-0.83
	(2.09)
Number of Polls	-0.21
	(.17)
Log(Ballots Cast)	-2.53*
	(1.21)
Number of Contests	74
R^2	0.33

The results are instructive. The average absolute horserace error in closed primaries is less than the average absolute horserace error in open primaries, but the differences are not statistically distinguishable from one another.[2] Caucuses are associated with much bigger errors – the average absolute horserace is nearly 10 points greater in caucuses and this is a statistically

[2] Mixed primaries are the omitted category.

significant difference. Relatedly, larger contests are associated with fewer polling errors – for every 1% increase in the size of the electorate the average absolute horserace decreases by 2.53% all else equal.

In general, it is hard to conclude that primary polls were noticeably worse that primary polls in earlier years, despite some high profile misses (e.g., polls in the Michigan Democratic primary). Moreover, while some states caused more trouble for pollsters than others, there are not many systematic features of either polls or contests that are related to the average accuracy of polls that lend much guidance going forward. Polls done further from Election Day contained more error, all else equal, as did polls predicting caucus outcomes. Polls seemed to do better when more polls were taken, but it is hard to know whether this reflects that polls were more likely to be conducted in some contests than others. While there will obviously always be outliers, and we have explicitly and intentionally avoided trying to estimate the impact of pollster-specific "house effects," the analyses reveal very little evidence that the ability of polls to predict the margin of victory systematically vary according to mode of interview, sample size, field period, or proximity to election day during the last two weeks.

What the results do suggest is a need for an increased sensitivity for the many errors that are present in pre-election polling. The 2016 primary polls did not perform noticeably worse than earlier primary elections, but there is a consistent level of error that is still more than twice the "margin of error" that polls publicly report. A heightened sensitivity to the errors involved in polling seems sensible going forward.

C.2 Error by Distance from Election Day

State polls that ended in the final 13 days were conducted slightly earlier than national polls, raising the possibility that state surveys failed to catch a late shift in Trump's direction. To

assess this, the distance between the middle of a poll's field period and Election Day was calculated for all battleground state and national surveys, allowing errors to be compared among earlier and later polls. The mid-date for state polls ending in the final 13 days averaged 7.8 days away from Election Day, while national polls averaged 6.4 days before the election's end.

National polls with a midpoint less than 5 days before the election (16) exhibited slightly higher errors than those conducted earlier in the final two weeks (2.0 vs. 1.6), and the average bias against Trump was apparent only in the final polls before the election (0.8 vs. -0.2).

State surveys with the midpoint less than 5 days before the election (3.6) as those conducted earlier in the final two weeks (3.7); the average bias underestimating Trump's support was slightly higher in polls completed closer to Election Day than earlier polls (2.6 vs. 2.3). While there was very little difference in accuracy using the five-day cut-off, the 22 state-level surveys with midpoints less than 3.5 days from the election proved more accurate. These surveys averaged a 2.7-point vote margin error and 1.4-point bias underestimating Trump, providing at some support for the theory that inaccuracy of state polls was due to a late shift in preferences.

C.3 Poll Performance during the 2016 Presidential Primaries

This section considers the accuracy of primary polls across the 2016 nomination timeline. Previous research indicates that performance during the primaries varies across states and particularly over time (Traugott and Wlezien 2009). What about 2016? Do we observe a similar pattern?

Little scholarship examines the accuracy of the polls during the nomination process. Beniger (1976) considered the relationship between the polls and primary outcomes from 1936 to 1972 and found that being the leader in early polls was the best predictor of electoral victory.

While not surprising, it is not clear what it tells us about the current nomination process, which emerged in 1972.

Only two pieces of research explicitly examine the performance of polls in the current nomination system – Bartels and Broh (1989) and Traugott and Wlezien (2009). Bartels and Broh analyzed the performance of three organizations (the CBS News/ *New York Times* poll, the Gallup Organization, and the Harris Poll) in the 1988 primaries, polling efforts during which were limited. Bartels and Broh also found inconsistencies in the reporting of the poll numbers. Still, Bartels and Broh made some observations, the most noteworthy of which is that the polls underestimated the support for each candidate (with the exception of Senator Robert Dole).

Two decades later, Traugott and Wlezien (2009) studied poll performance over the course of the 2008 nomination process. Their poll data came from published state-level results of public pollsters from the week preceding each primary or caucus – 258 polls in 36 different Democratic events and 219 polls in 26 Republican events – and their analysis focused on the gap between the winner candidate's vote share and poll share. They found that the vote share almost always exceeded the poll share while the race remained competitive, particularly early on in the nomination process. In an unusual perspective made possible by the length of the contest on the Democratic side in particular, this could be observed through most of the primaries; it was not the case in the Republican events after John McCain became the presumptive nominee. The analysis also shows there are state-specific contextual factors at work that can affect the quality of the estimates that public pollsters make.

Less directly relevant, though worth of note, is Hopkins' (2009) briefly study of a *Wilder effect* and *Whitman effect*—the tendency for voters to overestimate their support of African American candidates and underestimate their support of female candidates in statewide elections

for Governor and U.S. Senator across the period from 1989 to 2006. His analysis of general election polls found that there was a tendency to overstate support for African American candidates early in this period but that it disappeared after 1996, and polls never underestimated support for women. He extended his analysis to the 2008 Democratic primary series, looking specifically at the difference between poll support for Barack Obama and Hillary Rodham Clinton and their vote shares, and found that Obama consistently did slightly better in the elections than the polls suggested. This varied across states with the proportion of the black voters; the polls were generally accurate in primary states with few black voters but consistently understated Obama support in states with many black voters. This comports with what Traugott and Wlezien (2009) found and is the opposite of the "Wilder effect" that would have been predicted among white voters. Hopkins did not observe any "Whitman effect" for Clinton during the 2008 primaries.

The analysis relies on data identified for this report, and focuses entirely on published state-level results from the two weeks preceding each primary or caucus for which polls were available. This means that we do not have data for all states. All told, there are 457 polls, 210 of which relate to the 38 Democratic elections and 247 to 36 Republican events.[3] The polls that we do have also are not equal, as there is great variation in survey practices, including survey mode, question wording, likely voter modeling, weighting procedures, and sample size. This analysis does not attempt to take account of these differences, in part because of the difficulty of obtaining complete information. Other analysis in the report does address some of these issues,

[3] We do not have polls in the last two weeks for both the Democratic and Republican events in the following states: Alaska, Hawaii, Maine, Minnesota, Montana, Nebraska, New Mexico, North Dakota, South Dakota, Washington and Wyoming. Polls also are missing before for the Democratic primary in Kentucky and the Republican events in California, Colorado, and New Jersey.

and demonstrates fairly minimal effects. The poll estimates used in the analysis are simple averages of the results for each event. The specific variable of interest is the difference between the vote margin of the two leading candidates and the poll margin in the preceding two weeks:

(1^{st} place vote – 2^{nd} place vote) – (1^{st} place poll – 2^{nd} place poll).

Thus, the variable is positive when the winner outperforms the polls and negative when the winner underperforms, and it takes the value of "0" when the margins are equivalent. It is important to use a signed error term in place of the absolute error because this is informative about patterns of poll performance over time, as we will see.

We start with basic descriptive statistics of poll errors during 2016. Table C.2 summarizes means (and standard deviations) both for signed and absolute errors, first for all 74 primaries and caucus taken together and then for Democratic and Republican events taken separately. The signed errors in the first row indicate that the vote margin tended to exceed the poll margin across primaries and causes, by about 6.8 percentage points on average. This comports with the previous research, particularly Traugott and Wlezien (2009) but also Bartels and Broh (1989). The pattern was particularly pronounced for the Democrats, where the mean error in the vote-poll margin approached 9.6 points, by comparison with only 3.8 points in Republican events. The absolute errors in the second (main) row of Table C.2 reveal that this partisan "bias" in errors did not produce proportionately greater absolute error; indeed, the mean error for Democratic events was only 1.5 point higher on average, 13.1 vs. 11.6. That the polls performed about as well in absolute terms across the parties implies that signed errors tended to cancel out more for the Republicans than for the Democrats.

Table C.2 Primary Poll Performance in 2016: Mean Difference between Winner's Vote and Poll Margins

	All	Democrat	Republican

Signed error	6.8 (14.6)	9.6 (15.1)	3.8 (10.8)
Absolute error	12.4 (10.0)	13.1 (12.5)	11.6 (8.1)
n	76	38	36

Note – Standard deviations in parentheses.

Timing is not everything, of course. Poll performance can depend on other factors, including the level of support in the polls itself. That is, in states where a candidate is dominating in the polls, we might expect a very big lead to shrink. Traugott and Wlezien (2009) observed such a pattern in the 2008 primaries, and they also revealed that the poll margins themselves varied over time.[4] Table C.3 shows bivariate correlations between the timing of the primary, the difference between the vote and poll margins, and the poll margins themselves. The top part of the table contains results for all 74 primaries and caucuses. Here we see that the vote-poll margin is negatively related to the winner's poll margin, just as Traugott and Wlezien (2009) found. The error also is positively related to the number of days into the election year the primary occurs. The winner's poll margin itself does not appear to increase (or decrease) over the process.

Table C.3 Selected Correlates of Primary Poll Performance

	Winner's vote-poll margin	Winner's poll margin
All Primaries		
Winner's poll margin	-0.30 (.01)	–
Number of days into election year	0.20 (.09)	-0.03 (.79)
Democratic Primaries		
Winner's poll margin	-0.33 (.04)	–
Number of days into election year	0.02 (.88)	-0.28 (.09)
Republican Primaries		
Winner's poll margin	-0.26 (.13)	–

[4] That said, it is important to note that they focused on the winner's share of the top two candidates' poll shares in each primary.

Number of days into election year	0.42 (.01)	0.43 (.01)

Note – Two-tailed p-values in parentheses.

The overall set of results conceals differences between the parties. First, the vote-poll margin is negatively related to the winner's poll margin for both the Democrats and Republicans, though only significantly so for the Democrats. Second, the vote-poll margin is positively related to the primary date for both parties, though the relationship is strong and statistically significant only for the Republicans, much as we would expect given Figures C.5 and C.6. Third, the winner's poll margin also varies with the timing of the primary for both the Democrats and Republicans, though the relationship differs dramatically by party. That is the poll margin for the Democratic winner tended to decrease over time whereas the poll margin of the Republican winner (Trump) tends to increase. This difference may – at least in part – reflect the differences in the competitiveness of the race over time.

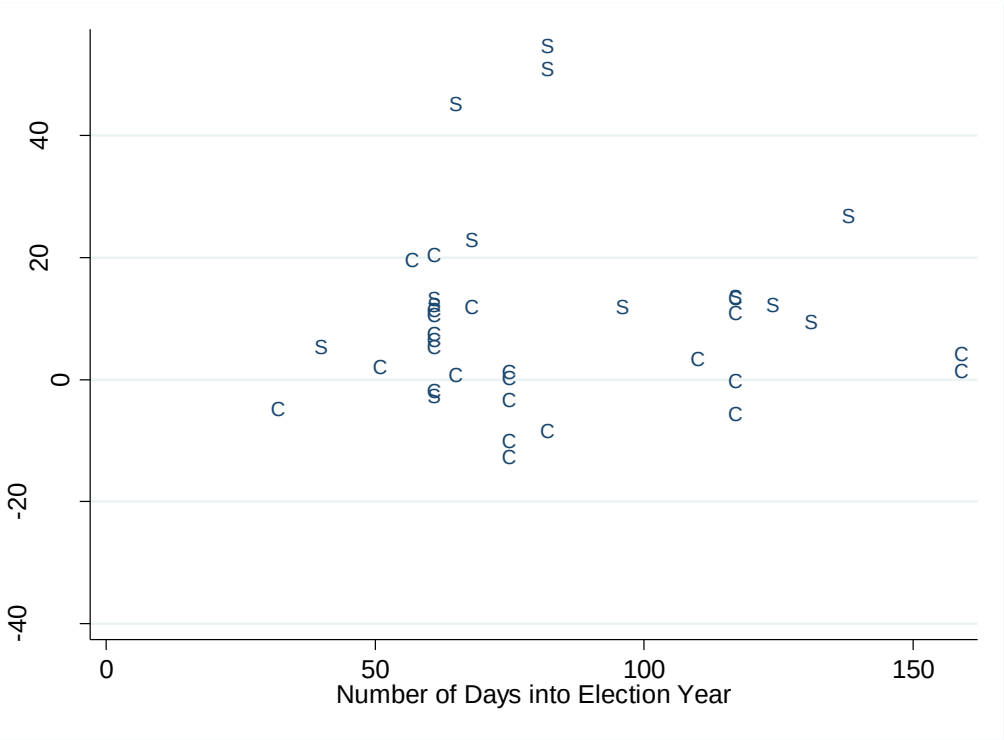

Figure C.5 The Polls and the Vote in the 2016 Democratic Presidential Primaries.
Notes – Each entry in the figure is the difference in a state between the winner's actual vote share and the share of the second place candidate minus the corresponding pre-election poll margin in the two weeks leading up to the election. A "C" indicates a Clinton win and an "S" represents a win by Sanders.

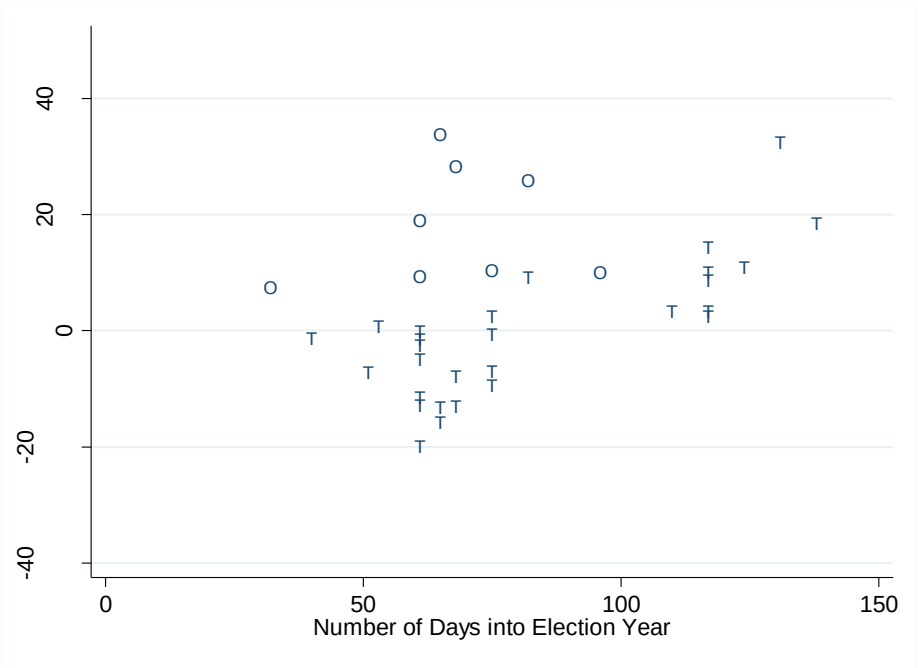

Figure C.6 The Polls and the Vote in the 2016 Republican Presidential Primaries.
Notes – Each entry in the figure is the difference in a state between the winner's actual vote share and the share of the second place candidate minus the corresponding pre-election poll margin in the two weeks leading up to the election. A "T" indicates a Trump victory and an "O" is used to represent a win by some other candidate.

The bivariate analyses are useful, but they only take us part of the way toward explaining the estimation errors in the pre-primary polls, and a multivariate analysis is required. Results of this analysis for the Democratic primaries are displayed in Table C.4. The first column contains results of a simple baseline regression containing the winner's poll margin. As expected given Table C.3, we see that poll leads have a significant negative impact on the vote-poll margin difference. The coefficient (-.26) should not be taken to imply that poll leads generally shrink, as we have already seen. Rather, the greater the poll margin, the less the winner's vote margin

exceeded it—for each additional four points in poll margin, the winner's vote-poll gap declines by one percentage point. With a poll share of about 50%, we predict no real difference between the vote and poll margins. With even larger shares, we would expect the poll margins to shrink by Election Day. The second column of Table C.4 adds the election timing variable. These results also are expected given what we have seen, as the campaign date just does not appear to matter for the vote-poll error in the Democratic primaries.

Table C.4 Regressions Predicting Winner's Vote Margin Minus Poll Margin, Democratic Primaries

	Model 1 (Baseline)	Model 2
Winner's poll margin	-0.26* (0.12)	-0.28* (0.13)
Number of days into election year	–	-0.04 (0.08)
Intercept	13.53* (2.53)	16.76* (7.72)
R-squared	0.11	0.12
Adj. R-squared	0.09	0.07
Root MSE	14.43	14.6

Notes – N = 38; * p < .05 (two-tailed)

Table C.5 shows a slightly different structure on the Republican side. In the first column, the winner's poll margin has a negative significant effect on error, and with virtually the same coefficient that we saw for the Democrats (-0.27 vs -0.26). In the second column, we can see the strong association noted earlier between the election date and the vote-poll margin. Indeed, the coefficient implies that we expect the signed error to increase by one-third of a percentage point each day of the nomination process. Given the intercept (-13.02), the result implies that the signed error would tend toward 0 through mid-February and then become increasingly positive, much as we saw in Figure C.6. When including the campaign date, the effect of winner's poll share doubles in size and easily exceeds even stringent levels of statistical significance. Based

on these results, the errors in polls varied in fairly predictable ways in the 2016 nomination process, particularly the Republican contests.[5]

Table C.5 Regressions Predicting Winner's Vote Margin Minus Poll Margin, Republican Primaries

	Model 1 (Baseline)	Model 2
Winner's poll margin	-0.27 (0.18)	-0.58* (0.16)
Number of days into election year	–	0.33* (0.07)
Intercept	8.00* (3.57)	-13.02* (5.55)
R-squared	0.07	0.41
Adj. R-squared	0.04	0.38
Root MSE	13.28	13.68

Notes – N = 36; * p < .05 (two-tailed)

Though polling misses in primary elections may be the rule more than the exception, there is a good amount of pattern to the errors we observed in 2016. To begin with, we see that the polls tended to underestimate the winner's vote margins. This tendency varies across candidates, being much more pronounced for insurgents, particularly early in the process. More generally, the performance tended to vary across space and time. The larger the poll lead in a particular state, the less the vote margin exceeds the poll margin, and timing also mattered, at least for the Republicans. Other features of context might matter as well, and separate analysis suggests that the black population of a state positively influenced Clinton's vote margin given the poll margin. (This parallels what Hopkins (2009) and Traugott and Wlezien (2009) found for Obama in 2008.) No such patterns were observed on the Republican side. While there are differences in the details, the general lesson is clear: poll performance in primary polls is different from what

[5] Analysis incorporating an interaction between number of days and the winner's poll margin significantly improves the fit of the model and increases the estimated effect of that margin, but indicate that its impact may decrease over time.

we observe in general elections and that performance itself varies across in understandable ways across the electoral calendar, the level of support in each state, and the specific characteristics of the state as well.

Appendix D. Approaches to Likely Voter Modeling

The assumptions that pollsters make about turnout and the methods they use to measure and model the likely electorate vary widely. More than a quarter century after Irving Crespi (1988) described identifying likely voters as "a major measurement problem in pre-election polling," this aspect of survey design remains a combination of science and art, with few pollsters taking the same approach. While a complete assessment of the various pollster likely voter models is beyond the scope of this report, we can summarize some of the most common approaches taken. Some pollsters make direct assumptions about the demographic and geographic composition of the likely electorate, and apply quotas or weights (or, more formally, pre or post-stratification) to assure that their final samples match these assumptions. One pollster, for example, weighted their Pennsylvania poll "to match expected turnout demographics for the 2016 General Election." While easier to explain and understand, this relatively direct approach is not the most typical.

More often, the assumptions that pollsters make about turnout are not about voter demographics directly, but rather about the *techniques* and *mechanisms* they use to select, screen for or otherwise model the likely electorate. The voter demographics that result are more a byproduct of their respective approaches than some deliberate and explicit set of assumptions. Again, the specific techniques vary widely. Many begin by attempting to interview a random sample of all adults. They will weight their full adult sample to match the known demographics of the adult population as measured by U.S. Census. They will then use some mechanism to select or model the "likely voters" from within their sample of all adults, and allow their demographics to vary without additional weighting.

This selection process can be a straightforward screen based on the answers to one or more survey questions, or it can be based on an index constructed from as many as seven or eight questions with a cut-off between likely and unlikely voters made at some level of the index. Some attempt to calibrate their cut-off point to some "assumption" about the *level* of coming turnout. Pollsters will select a smaller fraction of their sample of adults as likely voters if they expect a lower turnout, and a larger fraction if their assumptions point to a bigger turnout.

Other pollsters screen for registered or likely voters during the interview, retaining no demographic information about the non-voters they screen out. For the purposes of weighting, such pollsters are far more likely to make direct assumptions about the demographics of the electorate since they cannot weight to match all adults. Some will weight to match the geographic distribution of likely voters based on previous vote counts at the county or town level (on the theory that such data is both readily available and precise), but not weight or adjust the demographics of selected likely voters (on the theory that benchmarks of past demographics are often conflicting and less reliable).

Pollsters who weight to match "expected" demographics often differ in the sources they use to set their weighting targets, drawing variously from past exit polls, the CPS Voting Supplement surveys, estimates drawn from official "voter file" records or some combination of the three.

Some pollsters sample directly from voter files as a means of more accurately selecting likely voters, by restricting potential respondents to those actually registered to vote or with some past history of voting. Among pollsters who use RBS, some may only use the list to identify the *households* of registered voters, using survey questions and random methods to select a "likely voter" within each household. Others may request a *specific voter by name*, with

that person selected based on their prior history of voting, sometimes determined from a complex statistical model.

In recent years, some pollsters have moved to increasingly more advanced and complex efforts to model the likely electorate. These include the so-called “analytics” surveys, which leverage techniques like multiple regression and poststratification (MRP). Examples include YouGov (Rivers and Lauderdale 2016), Morning Consult (2016) and the approach used by Corbett-Davies, Gelman and Rothschild to model a New York Times Upshot survey in Florida (Cohn 2016). Again, this listing just covers some of the more prominent features in the methods used to model likely voters. Examine the methods used by any one pollster, and you will likely find combinations of the approaches listed above, where the explicit assumptions range from relatively scant and hands-off to heavy and highly complex.

Appendix E. Testing for *Shy Trump* Mode Effects in National Polls

This section presents details of analyses testing the *Shy Trump* hypothesis by examining differences in estimated candidate support between self- and interviewer administered polls – while attempting to control for other poll characteristics. Some characteristics that differentiate the polls are the number of days in the field, the use of a likely voter model and whether the poll is a tracking poll or not. The number of days has been considered as an indicator of higher response rates and quality (Lau, 1994). The use of a likely voter model – instead of using estimates based on registered voters – is thought to lead to better estimates, given the socio-demographical determinants of turnout; finally, tracking polls use small samples every day and publish moving average estimates. The generally small size of daily samples may have an impact on these estimates.

Table E.1. Profile of Polls by Mode of Administration

	Total	Live phone	Web	IVR/Online
Number of days in field	4.2	4.5	4.2	2.9
Use of LV model	93%	97%	89%	94%
Prop. tracking	31%	13%	37%	61%
Prop. nondisclosure	6.6	4.3	8.5	5.6

As shown in Table E.1, these characteristics are related to modes of administration. Among the 160 polls conducted during the period under study, the average number of days in the field is 4.5 for the live phone, 4.2 for the online polls and 2.9 for the IVR + Internet polls. In addition, the incidence of tracking polls varied widely by mode from 13 percent of live phone poll to 61 percent of IVR + Internet polls.[6] Finally, more than 90 percent of the polls used likely voter, and

[6] Notice that the tracking polls are entered in the data base only once during their period in order to avoid any dependency in the data.

there was no difference between modes on this factor.[7] Table E.2 shows the impact of change over time and of the design features on the estimates of support for Trump over the two main candidates and of support for all the candidates. The sample of 160 polls is reduced to 156 because of some missing information for four polls. Table E.2 shows that the change in voting intention during the period can be best portrayed using a cubic model, at least in the case of support for Trump and for the third party candidates. Support for Clinton follows a quadratic curve (an inversed U). This change over time explains around 13% to 15% of the variation in the estimates.

Whatever the estimate used, the use of a likely voter model is not related to the estimates of support[8]. However, the number of days in the field is related positively to estimates of support for both Trump (+.38 per day) and Clinton (+23 per day) and negatively to support for the third party candidates (-.62), which means that polls with longer interviewing periods tended to record less support for third party candidates. Since support for these candidates tended to be too high relative to the vote, the results are in line with the idea that longer field periods indicate better methodology.

[7] When pollsters published two types of estimates, only the likely voter estimate was retained in this analysis. Therefore, the analysis performed here does not compare likely voter estimates and registered voters estimates for the same polls but for different polls usually conducted by different pollsters.

[8] This is congruent with Blumenthal, Cohen, Clinton and Lapinsky (2016) who showed little difference between likely voters and registered voters.

Table E.2. Methods and Support for the Candidates from September 1st to Election Day

	Two main candidates	All candidates		
	Trump	Trump	Clinton	Other candidates
Intercept	48.7 ***	41.4 ***	43.8 ***	14.8 ***
Time variables				
Time	-0.09 ***	-0.09 ***	0.06 **	0.03
Time squared	0.00 **	0.00 *	0.00 **	0.00
Time cubic	0.00 ***	0.00 ***	0.00	0.00 ***
Explained variance	15.2%	13.1%	15.2%	12.9%
Methods variable				
Days in field	0.09	0.38 ***	0.23 *	-0.62 ***
Used LV model	-0.64	-0.59	0.60	-0.03
Tracking poll	0.82 *	0.98 *	-0.42	-0.56
Live phone	-1.76 ***	-0.84	2.35 ***	-1.51 *
Online poll	-2.04 ***	-2.52 ***	1.17	1.36
Explained variance	27.4%	26.5%	25.3%	24.5%
Non-disclosers	0.00	0.19 **	0.18 *	-0.37 ***
Explained variance	26.9%	30.2%	28.1%	32.9%
N	156	156	156	156

*: p<.05; **: p<.01; ***: p<.001

In addition, tracking polls estimate support for Trump more than 0.8 points higher than the other polls when estimating his support on the sum of the two main candidates, and almost one point higher, when we use the estimate of support for all the candidates. This higher estimate is split on the estimates of the other candidates.

Finally, the impact of mode, i.e., web and live phone compared to IVR + Internet, after controlling for change over time and the different methodological features, is somewhat more complex. The coefficients show that polls using live phone do show an estimation of Trump support over the two main candidates that is more than 1.7 points lower than IVR + Internet

polls' estimates. Web polls, however, also show a lower estimation of support for Trump, by more than two points.

The situation is somewhat different when we examine the impact of mode of administration on the support for all the candidates: Web polls' estimates for Trump are 2.5 points lower than IVR + Internet polls' estimates, but there are no significant difference between live phone polls and IVR + Internet polls. Analyses of support for Clinton and for the third party candidates show a significant difference between live phone estimates and IVR + Internet estimates of the support for Clinton (+2.3) and for the other candidates (-1.5). However, there are no difference between web polls and IVR + Internet polls for these candidates. In summary, Trump systematically fared worse in Web polls while Clinton fared better in live phone polls and the third party candidates in IVR + Internet polls.

We may therefore conclude that there is a difference between modes, but not one that would validate a *Shy Trump* hypothesis. For Trump, estimates differ mostly between the two types of self-administered polls while for the other candidates, the difference is between the interview and the self-administered modes. It is however possible that these differences according to mode are due to different causes, i.e., that the lower live phone estimates are due to *Shy Trump* supporters but the lower web estimates are due to other factors, like sampling for example.

Appendix F. Testing for *Shy Trump* Interviewer Effects in National Polls

This appendix presents estimated logistic regression models testing for effects from interviewer characteristics on expressed support for Trump, controlling for respondent demographics. While the effects associated with respondent demographics are highly significant in both the ABC News/Washington Post and Pew Research Center polls, the effects associated with interviewer gender and interviewer race are not. This analysis, thus, provides no evidence for the *Shy Trump* hypothesis.

Table F.1 Estimated Trump Support Regressed on Interviewer and Respondent Characteristics in Two National RDD Polls

	ABC News/Washington Post Tracking Poll			Pew Research Center October Survey		
	B	Sig.	S.E.	B	Sig.	S.E.
Interviewer Characteristics						
Male	0.01		0.054	-0.11		0.101
White	0.00		0.053	-0.07		0.099
Respondent Characteristics						
Male	0.40	***	0.052	0.33	***	0.097
White Non-Hispanic	0.93	***	0.076	0.87	***	0.154
Black Non-Hispanic	-2.09	***	0.180	-2.85	***	0.477
High school or less	0.78	***	0.064	0.93	***	0.125
Some college	0.70	***	0.062	0.58	***	0.115
(Intercept)	-1.42	***	0.089	-1.42	***	0.179
Cox & Snell R-Squared	0.131			0.172		
Sample size	6,818			2,008		

Notes: Regression models are logistic and not weighted. Spanish language cases are excluded from this analysis because they tend to be systematically assigned to non-white interviewers. Reference categories: interviewer gender (female), interviewer race (Non-white), respondent gender (female), respondent race (other), respondent education (college graduate).

References

Bartels, L. M., and Broh, C. A. (1989), "A Review: The 1988 Presidential Primaries," *Public Opinion Quarterly*, 53(4), 563-589.

Beniger, J. R. (1976), "Winning the Presidential Nomination: National Polls and State Primary Elections, 1936-1972," *Public Opinion Quarterly*, 40(1), 22-38.

Blumenthal, M., Cohen, J., Clinton, J., and Lapinsky, J. (2016), "Why The NBC News/ Survey Monkey Poll Now Tracks Likely Voters," NBC News, September 10, 2016.

Cohn, N. (2016), "We Gave Four Good Pollsters the Same Raw Data. They Had Four Different Results." *New York Times*, September 20, 2016. Retrieved from https://www.nytimes.com/interactive/2016/09/20/upshot/the-error-the-polling-world-rarely-talks-about.html?_r=0.

Crespi, I. (1988), *Sources of Accuracy and Error in Pre-Election Polling*. New York: Sage.

Hopkins, D. J. (2009), "No More Wilder Effect, Never a Whitman Effect: When and Why Polls Mislead about Black and Female Candidates," *The Journal of Politics*, 71(3), 769-781.

Lau, R. R. (1994), "An Analysis of the Accuracy of 'Trial Heat' Polls During the 1992 Presidential Election," *Public Opinion Quarterly*, 58(1), 2-20.

Morning Consult (2016), "How We Constructed Our 50-State Snapshot." Retrieved from https://morningconsult.com/2016/09/08/constructed-50-state-snapshot/.

Rivers, D., and Lauderdale, B. (2016). "The YouGov Model: The State of the 2016 Election," October 4, 2016. Retrieved from https://today.yougov.com/news/2016/10/04/YouGov-Model-State-of-2016/.

Traugott, M. W., and Wlezien, C. (2009), "The Dynamics of Poll Performance During the 2008 Presidential Nomination Contest," *Public Opinion Quarterly*, 73, 866-894.