AN EVALUATION OF 2016 ELECTION POLLS IN THE UNITED STATES

AD HOC COMMITTEE ON 2016 ELECTION POLLING

COURTNEY KENNEDY, Pew Research Center
MARK BLUMENTHAL, SurveyMonkey
SCOTT CLEMENT, Washington Post
JOSHUA D. CLINTON, Vanderbilt University
CLAIRE DURAND, University of Montreal
CHARLES FRANKLIN, Marquette University
KYLEY MCGEENEY, Pew Research Center[1]
LEE MIRINGOFF, Marist College
KRISTEN OLSON, University of Nebraska-Lincoln
DOUG RIVERS, Stanford University, YouGov
LYDIA SAAD, Gallup
EVANS WITT, Princeton Survey Research Associates
CHRIS WLEZIEN, University of Texas at Austin

The Committee was supported by the following researchers:
Junjie Chen, Andrew Engelhardt, Arnold Lau, Marc Trussler, Luis Patricio Pena Ibarra

[1] Several months after joining the committee, Kyley McGeeney took a position at PSB, her current employer.

EXECUTIVE SUMMARY

The 2016 presidential election was a jarring event for polling in the United States. Pre-election polls fueled high-profile predictions that Hillary Clinton's likelihood of winning the presidency was about 90 percent, with estimates ranging from 71 to over 99 percent. When Donald Trump was declared the winner of the presidency in the early hours of November 9th, it came as a shock even to his own pollsters (Jacobs and House 2016). There was (and continues to be) widespread consensus that the polls failed.

But did the polls fail? And if so why? Those are the central questions addressed in this report, which was commissioned by the American Association for Public Opinion Research (AAPOR). This report is the product of a committee convened in the Spring of 2016 with a threefold goal: evaluate the accuracy of 2016 pre-election polling for both the primaries and the general election, review variation by different survey methodologies, and identify significant differences between election surveys in 2016 and polling in prior election years. The committee is comprised of scholars of public opinion and survey methodology as well as election polling practitioners. Our main findings are as follows:

National polls were generally correct and accurate by historical standards. National polls were among the most accurate in estimating the popular vote since 1936. Collectively, they indicated that Clinton had about a 3 percentage point lead, and they were basically correct; she ultimately won the popular vote by 2 percentage points. Furthermore, the strong performance of national polls did not, as some have suggested, result from two large errors canceling (under-estimation of Trump support in heavily working class white states and over-estimation of his support in liberal-leaning states with sizable Hispanic populations).

State-level polls showed a competitive, uncertain contest… In the contest that actually mattered, the Electoral College, state-level polls showed a competitive race in which Clinton appeared to have a slim advantage. Eight states with more than a third of the electoral votes needed to win the presidency had polls showing a lead of three points or less (Trende 2016).[2] As Sean Trende noted, "The final RealClearPolitics Poll Averages in the battleground states had Clinton leading by the slimmest of margins in the Electoral College, 272-266." The polls on average indicated that Trump was one state away from winning the election.

…**but clearly under-estimated Trump's support in the Upper Midwest.** Polls showed Hillary Clinton leading, if narrowly, in Pennsylvania, Michigan and Wisconsin, which had voted Democratic for president six elections running. Those leads fed predictions that the Democratic *Blue Wall* would hold. Come Election Day, however, Trump edged out victories in all three.

There are a number of reasons as to why polls under-estimated support for Trump. The explanations for which we found the most evidence are:

- **Real change in vote preference during the final week or so of the campaign**. About 13 percent of voters in Wisconsin, Florida and Pennsylvania decided on their presidential vote

[2] In four of those battleground states, a subset, which included Florida and Pennsylvania, the average poll margin was less than a point – signaling that either candidate could win.

choice in the final week, according to the best available data. These voters broke for Trump by near 30 points in Wisconsin and by 17 points in Florida and Pennsylvania.

- **Adjusting for over-representation of college graduates was critical, but many polls did not do it.** In 2016 there was a strong correlation between education and presidential vote in key states. Voters with higher education levels were more likely to support Clinton. Furthermore, recent studies are clear that people with more formal education are significantly more likely to participate in surveys than those with less education. Many polls – especially at the state level – did not adjust their weights to correct for the over-representation of college graduates in their surveys, and the result was over-estimation of support for Clinton.
- **Some Trump voters who participated in pre-election polls did not reveal themselves as Trump voters until after the election, and they outnumbered late-revealing Clinton voters.** This finding could be attributable to either late deciding or misreporting (the so-called *Shy Trump* effect) in the pre-election polls. A number of other tests for the Shy Trump theory yielded no evidence to support it.

Less compelling evidence points to other factors that may have contributed to under-estimating Trump's support:

- **Change in turnout between 2012 and 2016 is also a likely culprit, but the best data sources for examining that have not yet been released.** In 2016, turnout nationwide typically grew more in heavily Republican counties than in heavily Democratic counties, relative to 2012. A number of polls were adjusted to align with turnout patterns from 2012. Based on what happened in 2016, this adjustment may have over-estimated turnout among, for example, African Americans, and under-estimated turnout among rural whites. Unfortunately, the best sources for a demographic profile of 2016 voters have either not been released or not been released in full. While we think this could have contributed to some polling errors, the analysis that we were able to conduct examining the impact of likely voter modeling shows generally small and inconsistent effects.
- **Ballot order effects may have played a role in some state contests, but they do not go far in explaining the polling errors**. State election rules led to Trump's name appearing above Clinton's on all ballots in several key states that Trump won narrowly (Michigan, Wisconsin and Florida). Being listed first can advantage a Presidential candidate by roughly one-third of one percentage point. Given that pollsters tend to randomize the order of candidate names across respondents rather than replicate how they are presented in the respondent's state, this could explain a small fraction of the under-estimation of support for Trump, but ballot order represents at best only a minor reason for polling problems.

There is no consistent partisan favoritism in recent U.S. polling. In 2016 national and state-level polls tended to under-estimate support for Trump, the Republican nominee. In 2000 and 2012, however, general election polls clearly tended to under-estimate support for the Democratic presidential candidates. The trend lines for both national polls and state-level polls show that – for any given election – whether the polls tend to miss in the Republican direction or the Democratic direction is tantamount to a coin flip.

About those predictions that Clinton was 90 percent likely to win... However well-intentioned these predictions may have been, they helped crystalize the belief that Clinton was a shoo-in for president, with unknown consequences for turnout. While a similar criticism can be leveled against polls – i.e., they can indicate an election is uncompetitive, perhaps reducing some people's motivation to vote – polls and forecasting models are not one and the same. As the late pollster Andrew Kohut once noted (2006), "I'm not a handicapper, I'm a measurer. There's a difference." Pollsters and astute poll reporters are often careful to describe their findings as a snapshot in time, measuring public opinion when they are fielded (e.g., Agiesta 2016; Easley 2016a; Forsberg 2016; Jacobson 2016; McCormick 2016; Narea 2016; Shashkevich 2016; Zukin 2015). Forecasting models do something different – they attempt to predict a future event. As the 2016 election proved, that can be a fraught exercise, and the net benefit to the country is unclear.

Presidential primary polls generally performed on par relative to past elections. The polling in the Republican and Democratic primaries was not perfect, but the misses were fairly normal in scope and magnitude. When polls did badly miss the mark, it tended to be in contests where Clinton or Trump finished runner-up. Errors were smaller when they finished first. This suggests that primary polls had a difficult time identifying wins by candidates other than the frontrunner.

A spotty year for election polls is not an indictment of all survey research or even all polling. The performance of election polls is not a good indicator of the quality of surveys in general for several reasons. Election polls are unique among surveys in that they not only have to field a representative sample of the public but they also have to correctly identify likely voters. The second task presents substantial challenges that most other surveys simply do not confront. A typical non-election poll has the luxury of being able to be adjusted to very accurate benchmarks for the demographic profile of the U.S. population. Election polls, by contrast, require educated estimates about the profile of the voting electorate. It is, therefore, a mistake to observe errors in an election such as 2016 that featured late movement and a somewhat unusual turnout pattern, and conclude that all polls are broken. Well-designed and rigorously executed surveys are still able to produce valuable, accurate information about the attitudes and experiences of the U.S. public.

A proposal for addressing the performance of state-level polling. As this report documents, the national polls in 2016 were quite accurate, while polls in key battleground states showed some large, problematic errors. It is a persistent frustration within polling and the larger survey research community that the profession is judged based on how these often under-budgeted state polls perform relative to the election outcome. The industry cannot realistically change how it is judged, but it can make an improvement to the polling landscape, at least in theory. AAPOR does not have the resources to finance a series of high quality state-level polls in presidential elections, but it might consider attempting to organize financing for such an effort. Errors in state polls like those observed in 2016 are not uncommon. With shrinking budgets at news outlets to finance polling, there is no reason to believe that this problem is going to fix itself. Collectively, well-resourced survey organizations might have enough common interest in financing some high quality state-level polls so as to reduce the likelihood of another black eye for the profession.

1. INTRODUCTION

Donald Trump's victory in the 2016 presidential election came as a shock to pollsters, political analysts, reporters and pundits, including those inside Trump's own campaign (Jacobs and House

2016). In the vast majority of U.S. presidential elections (95%), the winner of the national popular vote had also been the Electoral College winner (Gore 2016). That was not the case in 2016, when a divided result cast a critical spotlight on the polls' performance.

The national polls in 2016 indicated that Hillary Clinton would win the popular vote by about 3.2 percentage points. Taken together, those national polls were essentially accurate; the Democratic nominee ultimately won the popular vote by 2.1 points. In most presidential election years, one could reasonably conclude from this information alone that Trump's winning the presidency was unlikely and that the polls accurately measured Americans' vote preferences.

In the 51 contests that decided the presidency, Trump won 306 electoral votes and Clinton 232.[3] Looking at the polls at the state level, most states seemed firmly in the Republican camp or in the Democratic one. Pundits cited up to 13 battleground states where the campaigns suggested the race could be close. The polls in that group of states showed competitive races with Clinton apparently holding a consistent advantage in most. But the advantage was slim: eight states with a combined 107 electoral votes had average poll margins (%Trump-%Clinton) of three points or less (Trende 2016).

On the eve of the election, however, three types of information widely discussed in the news media pointed to a Clinton victory. All three turned out to be either misleading or wrong.

- The patterns in early voting in key states were described in numerous, high-profile news stories as favorable for Clinton, particularly in Florida and North Carolina (Silver 2017). Trump won both states.

- In the days leading up to November 8, several election forecasts from highly trained academics and data journalists declared that Clinton's likelihood of winning was about 90 percent, with estimates ranging from 71 to over 99 percent (Katz 2016).

- Polling data from the Upper Midwest showed Clinton leading, if narrowly, in Pennsylvania, Michigan and Wisconsin – states that had voted Democratic for president six elections running. This third deeply flawed set of data helped confirm the assumption that these states were Clinton's *Blue Wall* (e.g., Goldmacher 2016; Donovan 2016). On Election Day, Trump eked out victories in all three. More than 13.8 million voted for president in those states, and Trump's combined margin of victory was 77,744 votes (0.56%) (Wasserman 2017).

The day after the election, there was a palpable mix of surprise and outrage directed towards the polling community, as many felt that the industry had seriously misled the country about who would win (e.g., Byers 2016; Cillizza 2016; Easley 2016b; Shepard 2016).

1.1 This Report

The 2016 U.S. presidential election poses questions that political scientists, sociologists and survey researchers will be studying for years, if not decades. This report, commissioned by the

[3] Neither candidate actually received exactly that number of electoral votes due to several electors voting for candidates other than one they were pledged to.

American Association for Public Opinion Research (AAPOR), seeks to address *only* the performance of polls in 2016. Readers looking for an explanation of why Donald Trump won should consult other sources. The reasons Trump won and the reasons the polls missed may be partially overlapping, but this report only attempts to address the latter.

This report is the product of a committee convened well before the election, in the spring of 2016, with the goal of summarizing the accuracy of 2016 pre-election polling (for both primaries and the general election), reviewing variation by different poll methodologies, and identifying differences from prior election years. That was an ambitious task before November 8. In the early morning hours of November 9, the task became substantially more complex and larger in scope as the committee felt obligated to also investigate why polls, particularly in the Upper Midwest, failed to adequately measure support for Trump.

The committee is composed of scholars of public opinion and survey methodology as well as election polling practitioners. While a number of members were active pollsters during the election, a good share of the academic members were not. This mix was designed to staff the committee both with professionals having access to large volumes of poll data they knew inside and out, and with independent scholars bringing perspectives free from apparent conflicts of interest. The report addresses the following questions:

- Was the accuracy of polling in 2016 noticeably different from past elections?
- How well did the polls measure vote preference in the 2016 general election?
- How well did the polls measure vote preference in the 2016 primaries and caucuses?
- Did the accuracy of polls clearly vary by how they were designed?
- Did polls, in general, under-estimate support for Trump and, if so, why?

Many different types of data were brought to bear on these issues. This information includes poll-level datasets in the public domain that summarize the difference between the poll estimates and the election outcomes and provide a few pieces of design information (pollster name, field dates, sample size, target population, and mode). For 2016 polls conducted close to Election Day, the committee supplemented those datasets with information about weighting, sample source (e.g., random digit dial [RDD] versus voter registration-based sample [RBS] for telephone surveys) and the share of interviews conducted with landlines versus cell phones, where applicable. Adding these design variables was done manually through searches of individual press releases, news stories, methodology reports and pollster websites. In many cases, design information about a poll was missing or unclear, in which case the committee contacted individual pollsters to obtain the information.

In all, the committee reached out to 46 different polling organizations. Half (23) responded to our requests. Those who did respond were generous with their time and information. Not surprisingly, none of the organizations that did not respond are members of AAPOR's Transparency Initiative and most do not have staff who are active in AAPOR.

Generally, noncooperation with the committee's requests did not have a noticeable impact on work, with one notable exception. Surveys conducted using interactive voice response (IVR), sometimes called *robopolls*, were scarce at the national level (just three pollsters used IVR), but

they represented a large share of polling at the state level, particularly in Wisconsin and Michigan. Those IVR pollsters did not respond to our requests for microdata. Thus, the committee was unable to analyze that microdata along with data from firms using other methods, which could have been informative about polling errors in those states.

Given the large number of pollsters active during the election, the volume of polling and the reality that all pollsters structure their microdatasets differently, the committee was selective in asking pollsters for microdata. Since provision of microdata is not required by the AAPOR Transparency Initiative, we are particularly grateful to ABC News, CNN, Michigan State University, Monmouth University, and University of Southern California/Los Angeles Times for joining in the scientific spirit of this investigation and providing microdata. We also thank the employers of committee members (Pew Research Center, Marquette University, SurveyMonkey, The Washington Post, and YouGov) for demonstrating this same commitment.[4]

In the sections of the report that follow, the sets of polls analyzed may differ by section (e.g., national versus state-level; final two weeks versus full campaign). While this may be distracting, each section features what, in our judgment, was the best data available to answer each specific research question. At the top of each section, we describe the data and provide the rationale for our choices.

There are several different metrics quantifying error in election poll estimates, but this report focuses on two simple measures that are easily compared to past elections. The first error measure is absolute error on the projected vote margin (or "absolute error"), which is computed as the absolute value of the margin (%Clinton-%Trump) in the poll minus the same margin (%Clinton-%Trump) in the certified vote. For example, if a poll showed Clinton leading Trump by 1 point and she won by 3 points, the absolute error would be ABS(1 – 3) = 2. This statistic is always positive, providing a sense of how much polls differed from the final vote margin but not indicating whether they missed more toward one candidate or another.

The other key metric is the signed error on the projected vote margin (or "signed error"), which is computed in the exact same manner as the absolute error but without taking the absolute value. This statistic can be positive or negative, with positive values indicating over-estimation of Clinton's support and negative values indicating over-estimation of Trump's support. In the example above, if Clinton led by 1 point in a poll and won by 3 points, the signed error would be -2 points. When averaging absolute error and signed error across multiple polls, the signed error is always lower than (or equal to) the absolute error since positive and negative values are averaged together. Neither measure should be confused with whether polls were within the margin of sampling error, a statistic that applies to individual candidate support estimates but not the vote margin.

1.2 Theories About Why Polls Under-estimated Support For Trump

Since Election Day, dozens of theories have been put forward by politicians, pundits, pollsters and many others as to why the polls missed in 2016. Many such theories have fallen by the wayside since the final official vote totals were tallied, showing Clinton with a narrow lead. In the end, the final vote came close to what the national polls found, at least in aggregate.

[4] A list of the microdatasets made available to the committee is provided in Appendix Table A.0.

As we discuss later, many polls did a reasonable job at the national level in the general election and at the state level in the presidential primaries. But many did not. Much of our analytical focus is on assessing errors in the general election polls, but some of the possible sources of the errors also apply potentially to polls in the primaries.

Here is a summary of the major types of potential errors that we investigate in this report.

1.2.1 Late Deciding

Both Trump and Clinton had historically poor favorability ratings. One possibility is that these negative evaluations made it difficult for some voters to decide whether to vote and, then, difficult to decide for whom to vote. Unhappy with their options, many voters may have waited until the final week or so before deciding, a set of last-minute changes that polls completed a week out from the election would not have detected. Perhaps this included those who broke late for Trump, as well as potential Clinton voters who decided not to vote because they concluded she was going to win.

1.2.2 Reporting Error (or the *Shy Trump* Hypothesis)

During the primaries and the general election, political observers speculated that voters who were supporting Trump were less likely to admit this stance to pollsters than those supporting Clinton. Trump's controversial statements could have made it uncomfortable for some respondents to disclose their support for him to an interviewer. Thus, Trump voters would be less likely to express their true intentions.

1.2.3 Nonresponse Bias

Response rates in telephone polls with live interviewers continue to decline, and response rates are even lower for other methodologies. Thus, there is a substantial potential that nonresponse bias could have kept a given poll from accurately matching the election results.

Generally, decisions about responding to a poll are not strongly related to partisanship (Pew Research Center 2012). Studies have also shown, however, that adults with lower educational levels (Battaglia, Frankel and Link 2008; Chang and Krosnick 2009; Link et al. 2008) and anti-government views (U.S. Census Bureau 2015) are less likely to take part in surveys. Given the anti-elite themes of the Trump campaign, Trump voters may have been more likely than other voters to refuse survey requests.

1.2.4 Modeling Error: Weighting or Likely Voter Models Were Faulty

Many pollsters adjust their raw results to population benchmarks because of variations in how willing various subgroups in the population are to participate in polls. Younger people are quite difficult to find and interview, as are those with lower levels of education. Adjusting or weighting the raw data to take into account these differences is often required. But some pollsters did not weight their data by education in 2016.

Another possible source of error is in the different likely voter models or screens used by pollsters. If these models do not accurately reflect who votes, it is unlikely that the poll results will match the election results. Generally, a poll result based on likely voters tends to be more

Republican than the same result based on all registered voters (e.g., Perry 1973; Pew Research Center 2009; Silver 2014). But that was not always the case in 2016, suggesting likely voter models may not have been working correctly.

Likewise, Nate Cohn (2016b) and others have argued that the voting electorate was never as diverse or educated as shown in exit poll data. Current Population Survey data and voter file analysis show a whiter, less-educated electorate than the exit polls. Thus, polls weighting to past exit poll parameters may have missed the mark in 2016.

1.2.5 Ballot Order Effects

Political methodologists have documented a small but non-trivial bias in favor of candidates listed first on election ballots (e.g., Ho and Imai 2008; Miller and Krosnick 1998; Pasek et al. 2014). This bias is a version of a primacy effect, which is the tendency for people to select options presented near the top of a list when the list is presented visually, as on a ballot. To cancel out this effect, pollsters typically rotate the order of the candidate names presented to respondents. Most state boards of elections, however, do not rotate the order of candidate names, but list the presidential candidates in the same order in every county and every precinct. In states like Michigan, Wisconsin and Florida where Trump was listed first on the ballot state-wide, this order effect could have slightly boosted his support in the election relative to the polls (BBC News 2017; Pasek 2016; Gelman 2017).

2. PERFORMANCE OF POLLS IN 2016 RELATIVE TO PRIOR ELECTIONS

In the aftermath of the general election, many declared 2016 a historically bad year for polling. A comprehensive, dispassionate analysis shows that while that was true of some state-level polling, it was not true of national polls nor was it true of primary season polls. Key findings with respect to the performance of polls in 2016 relative to prior elections are as follows:

- National general election polls were among the most accurate in estimating the popular vote margin in U.S. elections since 1936, with an average absolute error of 2.2 percentage

points and average signed error of 1.3 percentage points. They correctly projected that Clinton would win the national popular vote by a small but perceptible margin.

- State-level general election poll errors were much larger, with an average absolute error of 5.1 points and an average signed error of 3.0 points. This was exacerbated by state polls, which indicated the wrong winner in Pennsylvania, Michigan and Wisconsin – states that collectively were enough to tip the outcome of the Electoral College.

- In 2016, national and state-level polls, on average, tended to under-estimate support for Trump, the Republican nominee. In 2000 and 2012, however, general election polls clearly tended to under-estimate support for the Democratic presidential candidates. So, while it is common for polling errors to be somewhat correlated in any given election, there is no consistent partisan bias in the direction of poll errors from election to election.

- The 2016 presidential primary polls generally performed on par with past elections. The 2016 pre-election estimates in the Republican and Democratic primaries were not perfect, but the misses were fairly normal in scope and magnitude. The vast majority of primary polls predicted the right winner, with the predictions widely off the mark in only a few states.

2.1 The Performance of National General Election Polls[5]

National presidential polls in the 2016 general election were highly accurate by historical standards, resulting in small errors and correctly indicating Clinton had a national popular vote lead close to her 2.1 percentage-point margin in the certified vote tallies. In terms of the average of absolute value differences between each poll's Clinton-Trump margin and the certified national popular vote margin, the final national 2016 polls' average error was 2.2 percentage points off the actual vote margin. As shown in Figure 1, the 2016 national polls tended to be more accurate than 2012 national polls (2.9 points average absolute error) and roughly similar to polling in 2008 (1.8 points) and 2004 (2.1 points). The level of error in 2016 was less than half the average error in national polls since the advent of modern polling 1936 (4.4 points), and also lower than the average in elections since 1992 (2.7 points).

[5] This analysis includes polls that had a final field date falling within 13 days of Election Day (October 26th or later) and if their field period began by October 16th. National poll analysis includes only a polling firm's final estimate to ensure comparability with historical data. Analysis of state-level polls, by contrast, includes all polls completed within the final 13 days, including multiple surveys from the same firm in the same state. The exclusion of pre-final estimates from national polls results provides a clearer historical comparison to analyses by the National Council on Public Polls, which is the source of data from 1936-2012 and only includes final estimates.

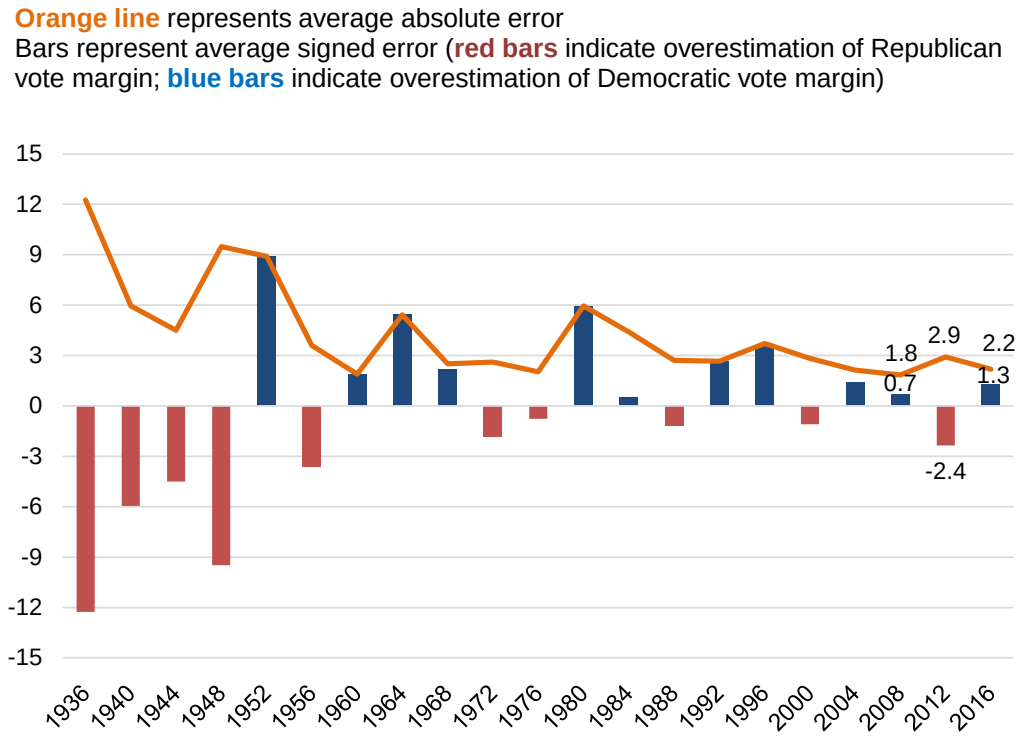


Figure 1. Average Error in Vote Margin in National Presidential Polls, 1936-2016.
Note – The 2016 figures are based on polls completed within 13 days of the election. Figures for prior years are from the National Council for Public Polls analysis of final poll estimates, some occurring before the 13-day period. Figures for 1936 to 1960 are based only on Gallup.

Examination of the average signed error in 2016 (1.3 percentage points) confirms that national polls in 2016 tended to under-estimate Trump's support more than Clinton's. The size and direction of error contrasts with 2012, when polls under-estimated Barack Obama's margin against Republican nominee Mitt Romney by 2.4 points. The average signed error in 2016 national polls was far lower than the typical level of signed error in either party's direction in presidential elections since 1936 (3.8 points), and is also lower than the 2.0-point average signed error in polls since 1992.

In recent elections, national polls have not consistently favored Republican or Democratic candidates. In 2016, national and state-level polls tended to under-estimate support for Trump, the Republican nominee. In 2000 and 2012, however, general election polls clearly tended to under-estimate support for the Democratic presidential candidates. Elections from 1936 to 1980 tended to show larger systematic errors and variation from election to election, in part, due to the small number of national polling firms.

Several media outlets combined national polls using varying methodologies to produce their own estimate of national support for Clinton and Trump, though none produced a more accurate estimate than the average of final national polls. RealClearPolitics estimated Clinton held a 3.2-point lead using a simple average of some final surveys, while FiveThirtyEight estimated Clinton held a 3.6-point margin in its "Polls-Only forecast" using a more complex method accounting for

systematic differences between pollsters' and historical accuracy. The Huffington Post estimated Clinton's lead at 4.9 percentage points nationally.

2.2 The Performance of State-Level General Election Polls

The trend line for state-level polls is similar to the trend line for national polls in one respect and very different in another. Unlike national polls, state-level polls in 2016 did have a historically bad year, at least within the recent history of the past four elections. Analysis of 423 state polls completed at least 13 days before the 2016 election, shows an average absolute error of 5.1 percentage points and a signed error of 3.0 percentage points in the direction of over-estimating support for Clinton. In the four prior presidential elections, the average absolute error in state polls ranged from 3.2 to 4.6.

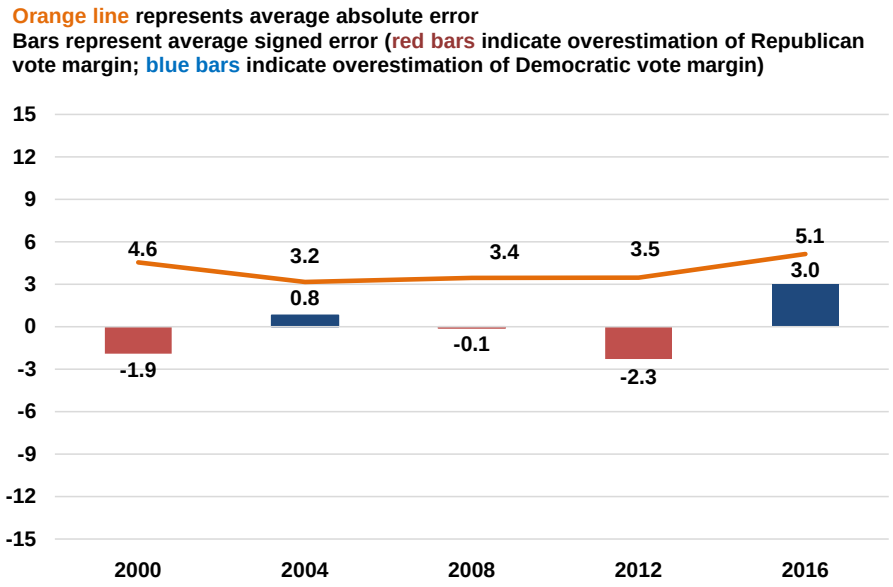


Figure 2. Average Error in Vote Margin in State Presidential Polls, 2000-2016.
Source – Figures for 2000 to 2012 computed from data made public by FiveThirtyEight.com.

The trend line for state polls is, however, similar to that for national polls in that there is no partisan bias. For a given election, whether the polls tend to miss in the Republican direction or the Democratic direction appears random. In 2016, the average signed error in state polls was 3 points, showing an over-estimation of support for the Democratic nominee. In 2000 and 2012, the average signed error in state polls was approximately 2 points, both times showing an over-estimation of support for the Republican nominee. While U.S. pollsters may be guilty of pointing to the wrong winner on occasion, as a group their work does not reveal any partisan leanings.[6]

[6] Even if U.S. pollsters collectively had a historical tendency to overestimate support for one major party relative to the other (e.g., as seen in U.K. polling since 1992 (Sturgis et al. 2016)), this would not be evidence of partisan behavior on behalf of the pollsters. Rather, such a pattern could be entirely (and more likely) explained by methodological factors, for example, over- or under-sampling African Americans or cell-phone only adults. We raise the issue of partisanship here only because that is one of the many criticisms leveled against the polling community from time to time. While the existence of a historical partisan error pattern would not prove partisan behavior, the absence of such a pattern should reassure poll consumers that pollsters are not putting their thumbs on the scale.

Both absolute errors and signed errors were smaller in battleground states, the 13 states that were decided by five points or fewer in the 2012 or 2016 presidential elections, than in non-battleground states (Appendix Table A.1). The average absolute error for the 207 battleground state polls was 3.6 points, compared with 6.4 points for the 206 polls in non-battleground states. The polls in non-battleground states under-estimated Trump's vote margin against Clinton by 3.3 points on average (signed error); the under-estimation of Trump's standing was 2.3 points in battleground states.

While the absolute errors tended to be lower in the more competitive states, under-estimation of support for Trump was substantial and problematic in several consequential states. Wisconsin polls exhibited the largest average signed error (6.5 points), with polls there showing Clinton ahead by between 2 and 12 points in the final two weeks before she narrowly lost the state (47.2 percent to 46.5 percent). Ohio polls also under-estimated Trump's margin by a substantial 5.2 points on average, indicating he had a small lead, though he went on to win the state by eight points. Polls in Minnesota, Pennsylvania and North Carolina each under-estimated Trump's margin against Clinton by an average of four to five percentage points, while polls in Michigan and New Hampshire under-estimated his standing by 3.5 percentage points on average. Under-estimation of support for Trump was smaller in Florida, Arizona and Georgia, while polls in Colorado and Nevada tended to over-estimate his support, and polls in Virginia exhibited little error.

2.3 The Performance of Primary and Caucus Polls

The 2016 presidential primary polls generally performed on par relative to past elections. The vast majority of primary polls predicted the right winner, with the predictions widely off the mark in only a few states. In short, the primary polls held their own in 2016. They improved in some important ways over previous years while retaining some weaknesses that the polling industry needs to note.

The committee based its analysis on all publicly released state-level candidate preference polls conducted in the final two weeks before each state's Republican and Democratic primaries. This totaled 457 state primary polls, including 212 polls in the Republican primaries and 245 polls in the Democratic primaries. Overall, there was at least one poll conducted in the last two weeks before the primary election in 78 of the contests. Additionally, the committee looked at the accuracy of the polling aggregator predictions made by three organizations: FiveThirtyEight, Huffington Post and RealClearPolitics.

Examining the polling averages in each state, the polls correctly pointed to the winner in 86% of the 78 primaries. This included correct predictions in 83% of the Democratic contests and 88% of the Republican contests. The misses were in three Republican primaries (Idaho, Kansas, Oklahoma) and in six Democratic primaries (Indiana, Kansas, Michigan, Oklahoma, Oregon, Rhode Island).

The average absolute error across all 457 state primary polls reviewed was 9.3 points,[7] not dramatically different from the performance of primary polls in other recent elections. While the average absolute error was higher in 2016 than in the four prior elections, a higher percentage of

[7] Appendix A.A discusses how a poll's margin of error relates to the accuracy metrics used in this report.

primary polls predicted the winning candidate in 2016 than was the case in 2008 and 2012. Analysis of the distribution of the size of primary poll errors in these recent elections (Appendix Figure A.1) shows a fairly stable pattern, with errors in 2016 polls looking similar to those in polls from other years.

Table 1. Performance of Presidential Primary Polls by Year

	2000	2004	2008	2012	2016
% Polls Predicting Winner	99%	100%	79%	64%	86%
Average Absolute Error	7.7	7.0	7.6	8.3	9.3
Number of Polls	172	129	555	195	457

2.4 Differences in General Election Poll Accuracy by Survey Design

A hallmark of the current election polling era is the tremendous variation in how polls are designed and conducted. Design variation is highly relevant to an examination of poll performance because survey researchers have long recognized that some approaches for constructing election polls are more accurate than others (Mosteller et al. 1949).

Many pollsters continue to use live telephone interviewing with random digit dial (RDD) samples of all the landlines and cell phones in the U.S. An even larger group conducts their surveys online, typically using opt-in samples of internet users. A third common approach is interactive voice response (IVR) either alone or in combination with an online opt-in sample. That combination is popular because IVR is only legal when dialing landline numbers, and so pollsters pair that with an opt-in internet sample in order to reach individuals who do not have a landline.

Nearly all IVR samples and an increasing number of live telephone samples are being drawn not from the RDD frames of all telephone numbers but instead from state-based voter registration files ("registration-based sampling," or RBS). While campaign pollsters have been using RBS for some time, the widespread use of RBS is a fairly recent development in public polls (Cohn 2014).

We examined two main design features for their effects on accuracy: mode of administration (e.g., live phone, internet or IVR) and sample source (e.g., RDD, RBS or opt-in internet users). We coded these variables for all national pre-election surveys and battleground state surveys conducted in the final 13 days of the general election. The data are summarized in Figure 3. While this typology does not encompass every final poll in 2016,[8] over 95 percent of the final two week polls conducted fall into one of these categories. Most IVR samples were selected using RBS, but in some cases the sample source was ambiguous. This is why the figures in this section do not attempt to make that distinction.

[8] Notable examples of other 2016 pre-election poll designs include internet surveys with a panel recruited offline using a probability-based sample (e.g., the USC Dornsife/LA Times Daybreak Poll) and mail surveys (the Columbus Dispatch Poll).

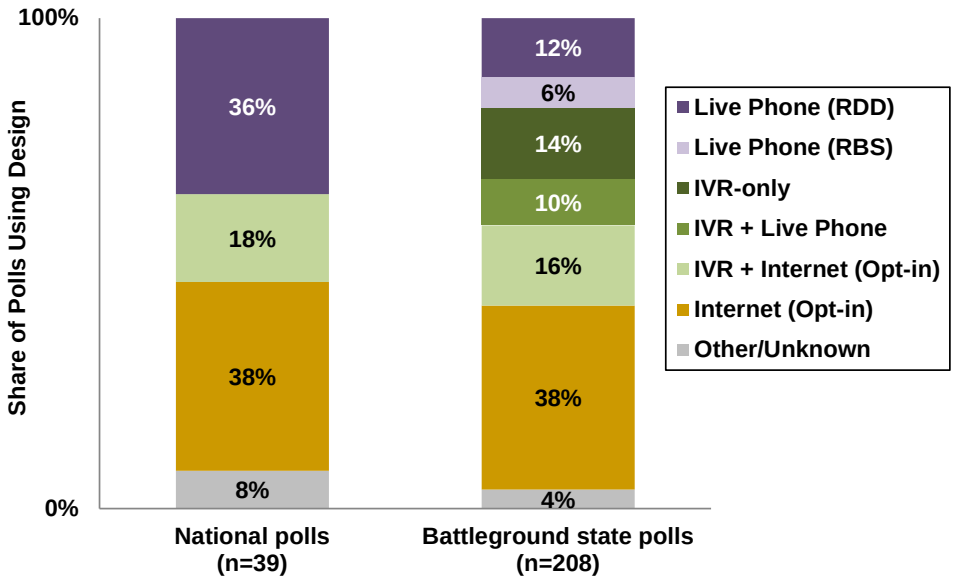


Figure 3. Design of 2016 General Election Polls Conducted in Final 13 Days.
Notes – The Franklin Pierce and Data Orbital polls, which were conducted by live telephone and had ambiguous statements about sample source that suggested RDD (but were not totally clear), are coded as live phone(RDD).

Several differences between national and battleground state polls are worth mentioning. In terms of mode, national polls were twice as likely to be conducted by live telephone as battleground state polls (36% versus 18%, respectively). Battleground state polls were about twice as likely to be conducted using some form of IVR as national polls (40% versus 18%, respectively). The share of polls conducted using the internet was basically the same for national and state-level polling.

Figure 4 gets to the central question of whether polls with certain types of designs were more accurate than others. Samples sizes for this analysis are small, and the effects from mode and sample source are to some extent confounded with house effects, such as differences in the likely voter model used. Still, IVR polls tended to exhibit somewhat less error in the 2016 general election than live telephone or internet polls. Battleground state polls that just used IVR had an average absolute error of 2.7 percentage points. By contrast, battleground state polls conducted using RDD with live phone and online opt-in had average errors of 3.8 and 3.9 points, respectively. Among national polls, none was conducted using just IVR. The national polls conducted by IVR and supplemented with an online sample had an average absolute error of 1.2 points, as compared with 1.6 for live telephone and 1.5 for online opt-in polls.

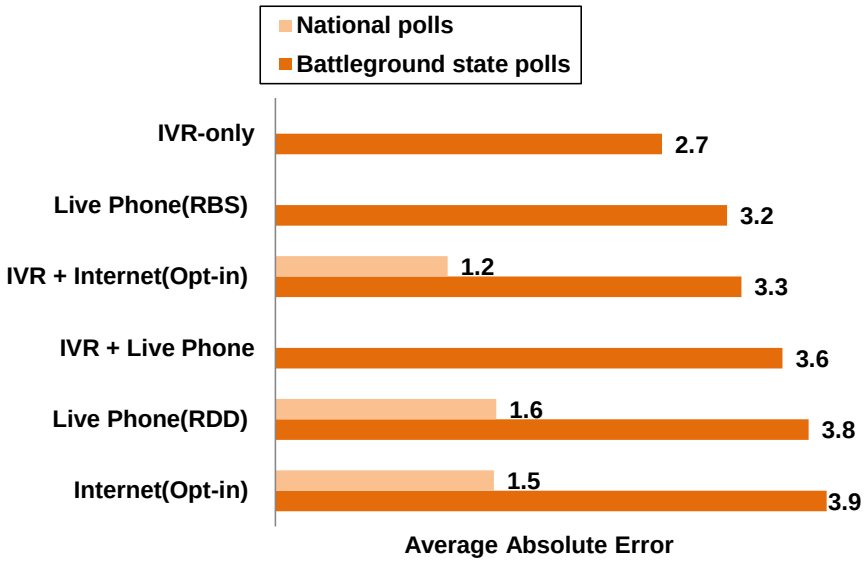


Figure 4. Average Absolute Error for 2016 General Election Polls, by Design.
Notes – Figures based on polls conducted during the final 13 days. Samples sizes for this analysis are small, and the effects from mode and sample source are to some extent confounded with house effects. National poll averages are based on 7 polls (IVR+internet), 14 polls (live phone RDD) and 15 polls (internet opt-in). Battleground state poll averages are based on 30 polls (IVR), 25 (live phone RBS), 34 polls (IVR+internet), 20 polls (IVR+live phone), 25 polls (live phone RDD) and 78 polls (internet opt-in).

In one respect, the fact that IVR-only polls did relatively well is surprising because federal laws dictate that IVR can only be used with landline numbers and about half of adults do not have landlines (Blumberg and Luke 2016). This half of the population would not have any chance of selection in an IVR sample assuming that cell phone numbers were flagged and purged before the IVR dialing began. Such substantial noncoverage usually increases the risk of bias.[9]

On the other hand, adults who have dropped their landline in favor of a cell phone or never had a landline to begin with tend to be younger and more racially and ethnically diverse than adults accessible by landline. These cell-only adults are more likely to be Democratic. In the 2016 election, in which turnout among African Americans and younger voters was not particularly high, under-coverage of cell phone-only voters appears not to have been a major problem and may help explain why IVR-only polls performed relatively well. In fact, when IVR polls were supplemented with an online component to capture cell phone-only voters, they did slightly worse.

Analysis of national polling errors by mode in recent elections (Figure 5) shows that IVR polls did not do particularly well in 2008 and were only nominally better in 2012 – elections in which Democratic turnout was relatively high. In fact, internet polls fared the best in both 2008 and

[9] One notable lesson from reviewing countless methodological reports is that the commonly held assumption that IVR polls just dial landlines (Cassino 2016; Clement 2016; Clinton and Rogers 2013; Cohn 2014; Enten 2012; Jackson 2016; Pew Research Center 2016) is not always correct. At least two pollsters clearly described their methodology as just IVR and yet reported that a noticeable share (10 to 25 percent) of their completed interviews were with cell phones. This detail may help explain why coverage error (e.g., excluding the half of the US population that is cell phone-only) may not have been more of a problem for IVR-only polls in 2016.

2012, with live phone polls in the middle. This indicates that the IVR results in 2016 are likely an election-specific phenomenon related to the particular turnout patterns that year.

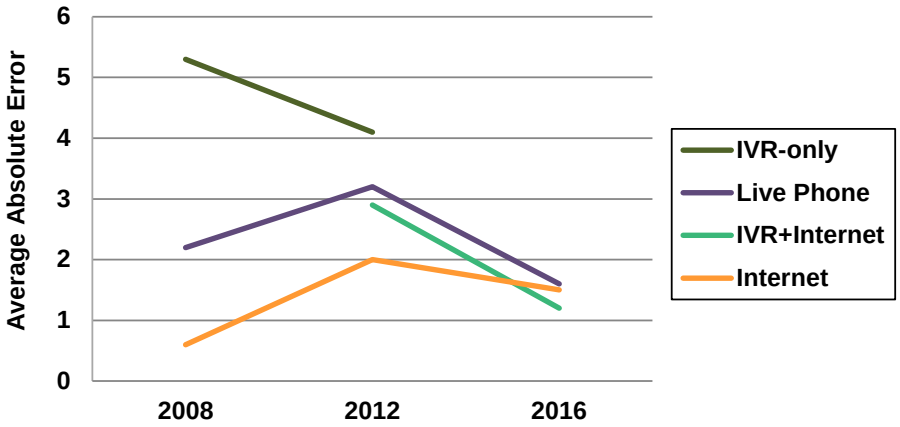


Figure 5. Absolute Average Error in National Polls by Mode by Year.
Note – In 2016 there were no national polls conducted using only IVR.

While the bivariate analysis presented in Figure 4 provides a high-level look at how accuracy varied by mode and sample source, it has a number of limitations. The number of polls and pollsters using each design during the final 13 days is modest at best and assignment to a given feature is not at all random. For example, polls using IVR were more likely than other types of polls to be conducted by partisan pollsters, especially Republican-affiliated pollsters. This raises the possibility that the relatively good performance on IVR polls in 2016 may have been due in part to some Republican pollsters making turnout assumptions slightly more favorable to Republicans. How much of the accuracy should actually be attributable to the IVR methodology per se is unclear.

The varying difficulty in predicting battleground state outcomes and the fact that some polls were fielded closer to Election Day than others can affect bivariate comparisons of accuracy. To better isolate the impact of methodological features in the polls, two ordinary least squares regression analyses examined the association of absolute error with mode and sample source, controlling for the geography in which the poll was conducted and the number of days between the election and the middle date of its field period. The results are reported in Appendix Table A.2.

The first regression model testing the association between mode and accuracy found that after taking geography and the number of days from the election into account, usage of IVR methods alone were associated with roughly 1-point lower absolute error than live-interviewer surveys, while internet, IVR/cell and IVR/internet polls did not have significantly larger or smaller errors than those conducted with live telephone interviewers. Usage of other, less common modes was associated with greater errors.

The second model focusing on sample source found no significant association of different sample sources with absolute error in vote margin estimates compared with RDD samples, when

state and timing are taken into account. Both regression analysis confirmed battleground state polls exhibited greater errors than national polls, particularly in Wisconsin, New Hampshire, North Carolina, Minnesota, Ohio, Pennsylvania and Michigan.

In sum, the regression analysis confirms the bivariate result that polls using IVR tended to have less error in the 2016 general election. It also indicates that mode was a more important predictor of error than sample source. This also suggests that the fact that IVR polls nearly always used voter file sample is perhaps not the sole or even the dominant reason for their relatively good performance. Taken together with the bivariate results, including those from past elections, it appears the accuracy of IVR polls may be a 2016-specific phenomenon: live telephone and internet polls did better in the recent past and may surpass IVR once again in a future election with different turnout patterns.

2.5 Difference in Primary Poll Accuracy by Design Features and Contest Features

Turning to the primary polls, regression analysis was also used to evaluate the effects of different design features on the accuracy of these polls. The model, presented in full in Appendix A.B, yielded the following main findings:[10]

- Survey mode had little effect**.** The differences in the absolute error of surveys employing different interviewing modes were not statistically significant. While polls using IVR and online methods are associated with slightly larger average absolute errors than polls with live interviewers, all else equal, the differences are small (0.21 and 0.08 larger, respectively, than a telephone poll) and not statistically distinguishable from zero.

- Caucuses were problematic. Caucuses were associated with much bigger poll errors than primaries. The average absolute error was nearly 10 points greater in caucuses – a statistically significant difference.

- Type of primary did not matter. There was no statistically significant difference in the accuracy of polls conducted in open vs. closed primaries.

- Size of the electorate (population) mattered. Larger contests were associated with fewer polling errors. For every 1% increase in the size of the electorate the average absolute error decreased by 2.5%, all else being equal.

- Certain states were harder to poll in than others. After holding all other factors constant, polls in Utah, South Carolina, Oregon, Michigan and Kansas were still off by a significantly greater margin than polls in other states. While it is impossible to diagnose the exact reasons for these systematic errors, controlling for them in the analysis is

[10] A note of caution is in order. Not every survey mode (telephone, internet, IVR) was used for every primary, or in the same proportions. This lack of uniformity across the 78 contests means that evaluations of state-level accuracy could be affected by differences in the types of survey mode employed in the states. Likewise, evaluations of the accuracy of survey mode could be affected by the types of states where each mode was used. (For instance, 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 was harder to predict, this would reduce our ability to interpret these differences as reflecting the impact of survey mode.)

important because it removes the impact of these state-specific errors from the estimated effects of other factors.

- The winner's lead mattered. There was important variation in average poll performance depending on whether the election was a blowout or not. Errors tended to be larger in uncompetitive contests.

Another major player in the polling scene during the primaries were the aggregators. The committee examined a total of five different estimation methods produced by three polling aggregators:

1. FiveThirtyEight
    a. 538 Polling Average (simple weighted average of the polls)
    b. 538 Polls Only (primarily based on polls, with limited adjustments)
    c. 538 Polls Plus (combines polls with an economic index; makes certain adjustments for historical election patterns)
2. Huffington Post Pollster (poll-based time series model)
3. RealClearPolitics (simple unweighted average of polls)

There were no significant differences among the aggregators in their prediction accuracy in primary elections. Since RealClearPolitics is using a simple unweighted average of the polls, and there was no statistically significant difference in accuracy between this method and the others, this means that additional modeling did not greatly increase the accuracy of the predictions.

The average signed error in the margin across all 230 aggregator predictions was -4.7, indicating that the predictions under-estimated the margin by 4.7 percentage points. The absolute error was much greater. The average error across all of the aggregators was 8.3, indicating that the average difference between the margin calculated by the aggregators and the actual margin for the winner was 8.3 percentage points. Although there appear to be significant differences across the aggregators in the absolute error overall, this is explained by different aggregators making predictions in different races. When only the same set of states are examined, there are no significant differences across the aggregators in the average absolute error (an average of 7.3 percentage points). There was no significant difference in the signed or absolute error between the Democratic and Republican contests for the aggregators, either overall or for the more commonly-polled contests.

3. EVIDENCE FOR THEORIES ABOUT WHY POLLS UNDER-ESTIMATED TRUMP'S SUPPORT

3.1 Late Deciding

One of the central hypotheses about why polls tended to under-estimate support for Trump is late deciding. Substantial shares of voters disliked both major party candidates (Collins 2016; Yourish 2016) and may have waited until the final days before deciding. If voters who told pollsters in September or October that they were undecided or considering a third party candidate ultimately voted for Trump by a large margin, that would explain at least some of the discrepancy between the polls and the election outcome. There is evidence that this happened, not so much at the national level, but in key battleground states, particularly in the Upper Midwest.

As reported by Blake (2016), the National Election Pool (NEP) exit poll conducted by Edison Research showed substantial movement toward Trump in the final week of the campaign – particularly in the four states Clinton lost by the smallest margins. In Michigan, Wisconsin, Pennsylvania, and Florida, 11 to 15 percent of voters said that they finally decided for whom to vote in the presidential election in the last week. According to the exit poll, these voters broke for Trump by nearly 30 points in Wisconsin, by 17 points in Pennsylvania and Florida, and by 11 points in Michigan. If late deciders had split evenly in these states, the exit poll data suggest Clinton may have won both Florida and Wisconsin, although probably not Michigan or Pennsylvania, where Trump either won or tied among those deciding before the final week. This pattern was not nearly as strong nationally.

Table 2. Time of Decision and Presidential Vote in Key States Won by Trump

	% Voters who decided in final week	Vote choice among voters deciding in final week		Vote choice among voters deciding earlier		Estimated Trump gain from late deciders	Election (%Trump-%Clinton)
		Trump	Clinton	Trump	Clinton		
Florida	11%	55%	38%	48%	49%	2.0%	1.2%
Michigan	13%	50%	39%	48%	48%	1.4%	0.2%
Pennsylvania	15%	54%	37%	50%	48%	2.3%	1.2%
Wisconsin	14%	59%	30%	47%	49%	4.3%	0.8%
National	13%	45%	42%	46%	49%	0.8%	-2.1%

Note – Analysis from Aaron Blake (2016) using NEP exit poll data.

Overall, these exit poll data suggest that voter preferences moved noticeably, particularly in these four decisive states. This can be seen as good news for the polling industry. It suggests that many polls were probably fairly accurate *at the time they were conducted*. Clinton may very well have been tied, if not ahead, in at least three of these states (MI, WI, FL) roughly a week to two weeks out from Election Day. In that event, what was wrong with the polls was projection error (their ability to predict what would happened days or weeks later on November 8th), not some fundamental problem with their ability to measure public opinion.

3.1.1 Explanations for Late Decisionmaking

The notion that pre-election polls fielded closer to Election Day tend to be more predictive of the election outcome than equally rigorous polls conducted farther out is not only intuitive, it has also been well documented for some time (e.g., Crespi 1988; Traugott 2001). The effect of late changes in voters' decisions can be particularly large in elections with major campaign-related

events very close to Election Day (AAPOR 2009). The 2016 general election featured a number of high profile campaign-related stories, as summarized in Table 3. Perhaps the most controversial single event was the FBI director's announcement on October 28th that the agency would review new evidence in the email probe focused on Clinton. The Clinton campaign claimed that that event was decisive in dooming her electoral chances (Chozick 2016).

There were other major events that also could have changed a substantial number of voters' minds. The Access Hollywood video tape released October 7th seemed to noticeably affect the race (Bradner 2016; Salvanto 2016), but that occurred too far out from Election Day to explain the errors observed in polls conducted during the final week or two of the campaign. Other events, such as the circulation of fake news stories (e.g., see Kang 2016) and Russian interference in the election (Director of National Intelligence 2017), could have influenced voters' decisions but seemed to emerge over time rather than at a clearly-defined point in the election.

Table 3. Major Events in the 2016 Presidential Campaign Following the Conventions

Aug. 1	Trump criticizes gold star family
Aug. 10	Judicial Watch releases State Department emails related to Clinton Foundation
Sep. 9	Clinton makes "basket of deplorables" comment
Sep. 11	Clinton leaves 9/11 ceremony early due to illness
Sep. 26	First presidential debate
Oct. 1	NYT reports Trump's 1995 tax record suggests no federal taxes for years
Oct. 3	NY attorney general sends cease and desist letter to Trump Foundation
Oct. 4	Vice presidential debate
Oct. 7	Release of video of Trump discussing groping women
Oct. 7	WikiLeaks releases emails hacked from Clinton campaign manager John Podesta
Oct. 9	Second presidential debate preceded by surprise Trump press conference
Oct. 12	Multiple women accuse Trump of touching them inappropriately
Oct. 19	Third presidential debate
Oct. 25	Announcement that Obamacare premiums will increase 25% on average
Oct. 28	FBI Director James Comey announces review of new evidence in Clinton email probe
Nov. 6	FBI Director James Comey announces emails warrant no new action against Clinton
Nov. 8	Election day

With one-time events, one might reasonably interpret a subsequent change in the horserace as an effect from that event. With ongoing, diffuse news stories, by contrast, it is not clear how one could measure the impact with polling data alone. Even under the cleanest of circumstances (i.e., a one-time event with no major competing news stories), the absence of a counterfactual makes investigations into the effect of particular campaign events a fraught exercise. In a hypothetical scenario in which the event of interest did not occur, a change in the horserace might still have been observed, just for different reasons.

Still, given the volume of claims that the FBI announcement of October 28th tipped the race in Trump's favor, we felt it worthwhile to investigate whether there was support for that claim in the public polls. We examined the five national tracking polls conducted during the final three weeks of the campaign. The margins (%Clinton-%Trump) for these polls are plotted in Figure 6.

Unlike other sections of this report (which focus on polling error), the goal of this particular analysis was to track voter sentiment as accurately as possible. To that end, Figure 6 presents the best available estimates for each tracking poll, which in the case of two polls meant using estimates produced with revised weights that better adjusted for sample imbalances (Cho et al. 2016; Tedeschi 2016).

The trend lines of the tracking polls in the figure are not very consistent with one another. For example, the ABC News/Washington Post poll (blue line) shows Clinton's support dropping precipitously in late October then rebounding before Election Day. The IBD/TIPP poll (yellow line) suggest a contradictory pattern, in which support for Clinton increased modestly in late October then tapered off in November. To try to detect a signal among these five somewhat unharmonious tracking polls, we computed the average margin giving each poll equal influence. It is interesting to note that this average shows the exact result of the popular vote (Clinton +2), which provides some confidence that collectively these polls were doing a reasonable job tracking voter preferences during this final stretch.

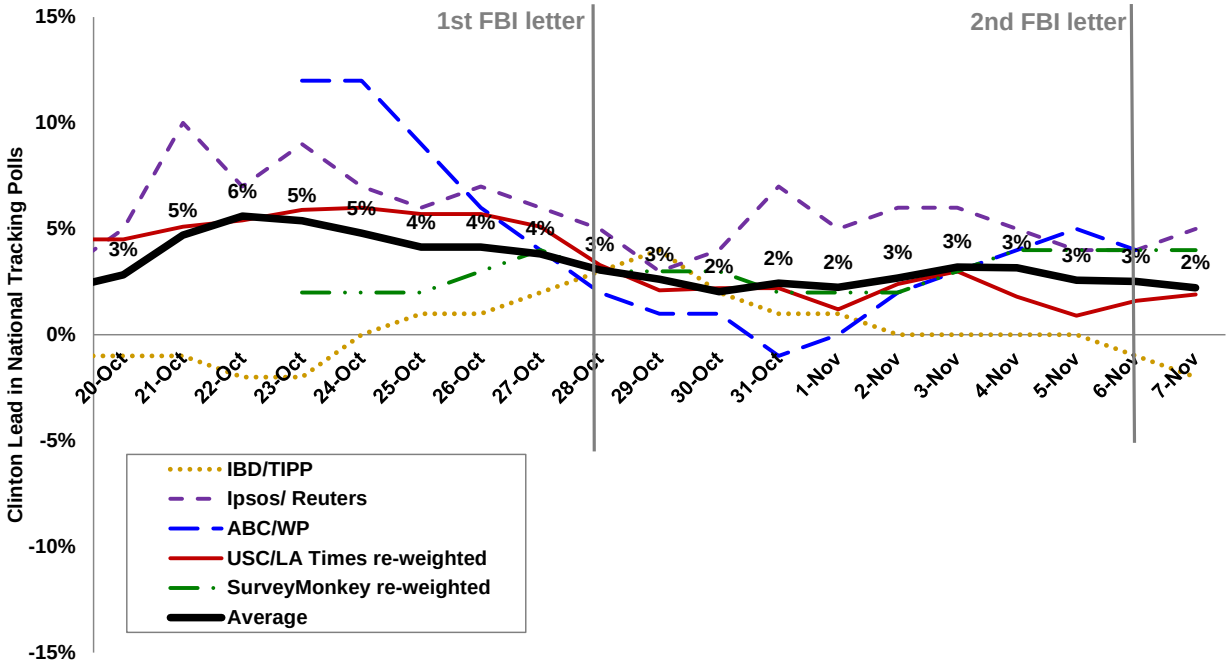


Figure 6. Presidential Vote Margin (%Clinton-%Trump) in National Tracking Polls.

The evidence for a meaningful effect on the election from the FBI letter is mixed at best. Based on Figure 6, it appears that Clinton's support started to drop on October 24th or 25th. October 28th falls at roughly the midpoint (not the start) of the slide in Clinton's support. What's more, the lag between when interviewing was conducted and when tracking poll results are released means that the slide in Clinton's support probably began earlier than estimates in Figure 6 suggest. For example, the ABC News/Washington Post estimate of a tied race on October 31 was based on interviews conducted October 28-31. The IBD/TIPP estimates are based on interviews

conducted during the six days prior to the date shown. Factoring in this lag, it is reasonable to speculate that Clinton's slide began as early as October 22 or 23. There were no notable campaign events on either of those days, though the announcement that Obamacare premiums will increase occurred roughly around that time (October 25th).

While Figure 6 indicates that Clinton's lead was eroding before October 28th, it is possible that the FBI letter news story made that erosion more severe than it otherwise would have been. Another way to analyze a possible impact of the first FBI letter is to check whether, all else equal, the trend in support changed following the release of that letter. To test this, we conducted a regression analysis using all national public polls fielded between September 1st and Election Day. This analysis, which controlled for change over time and methodological characteristics of the polls, indicates that the Comey letter had an immediate, negative impact for Clinton on the order of 2 percentage points. The apparent impact did not last, as support for Clinton tended to tick up in the days just prior to the election.

Based on all of the data examined here, we would conclude there is at best mixed evidence to suggest that the FBI announcement tipped the scales of the race. Pairing this analysis with the preceding one on NEP data for late deciders, it remains unclear exactly why late-deciding voters broke for Trump in the Upper Midwest. Anecdotal reporting offered a number of other suggestions (e.g., Republicans skeptical of Trump finally "coming home," Clinton's campaign – believing the Upper Midwest was locked up – allocating time and money elsewhere, Democrats lukewarm on Clinton deciding to stay home), but ultimately the data available do not offer a definitive answer to this question.

3.1.2 Callback Studies: Did Voters Tell Pollsters One Thing But Do Another?

If substantial shares of voters made up their mind about presidential vote very late in the campaign, one tool that should capture those late changes is a callback study. In a callback study the same people are interviewed before the election and after the election. Late change would manifest as discrepancies between pre- and post-election responses. It is also possible that *Shy Trump* responses would manifest the same way. Some poll respondents might have been inclined to censor their support for Trump before the election, but in light of his victory decide to be forthcoming about their vote for him in the post-election interview. So if poll respondents said in October that they were undecided and then said in November that they voted for Trump, the explanation could be either that they truly were undecided in October or that they intentionally misreported as undecided. For some voters, the truth may fall somewhere in between.

While callback data cannot necessarily distinguish between real late change and intentional misreporting, it can help to disentangle measurement error (which includes both *Shy Trump* answering and late switchers) from other error sources. Specifically, if a callback study shows that some respondents did not report being a Trump supporter before the election but nonetheless said they voted for him in the re-interview, that would indicate that measurement error was at least partially to blame for the poll's error rather than nonresponse bias (e.g., not enough Trump voters were in the study to begin with).

To test this, we examined data from the Pew Research Center's callback study. The study re-contacted registered voters in Pew's August and October national cross-sectional dual frame

RDD surveys. The re-interview was conducted by Princeton Survey Research Associates International November 10-14, 2016. Only respondents who self-reported having voted were eligible to complete the post-election re-interview (n=1,254). The crosstabulation of their pre-election and post-election responses are shown in Table 4.

Table 4. Comparing Individuals' Pre- and Post-election Responses to Presidential Vote

	Reported Vote				
Pre-election vote preference	Voted for Clinton	Voted for Trump	Voted for other candidate	DK or Refused	
Clinton/Lean Clinton	44.2%	0.4%	1.2%	0.6%	
Trump/Lean Trump	0.3%	38.2%	0.3%	1.1%	
Other candidate	1.6%	2.6%	6.3%	0.2%	
DK-Refused to Lean	0.7%	1.4%	0.4%	0.6%	
					100%

Source: Pew Research Center 2016 Election Callback Study. Based on 1,254 completed re-interviews with survey respondents who said that they voted in the general election. Estimates are unweighted.

Cases on the left-to-right diagonal represent respondents who answered the presidential vote question the same way before and after the election. About nine-in-ten respondents (89 percent) answered consistently while 11 percent reported something different at the ballot box than what they told the pollster before the election. In the context of recent elections, that 11 percent is quite typical. Pew Research Center has been conducting callback studies since 2000. Over the past five cycles, 12 percent of respondents, on average, were inconsistent in their pre- and post-election responses (i.e., were in an off-diagonal cell). The highest level of inconsistent responding recorded by Pew's callback studies was 18 percent in 2000, and the lowest was 7 percent in 2012.

What is notable about the 2016 data is not how many inconsistent respondents there were, it is how the inconsistent responders voted. Figure 7 shows the presidential vote margin among respondents who gave inconsistent pre- versus post-election responses in each callback study since 2000. Typically, those who admit changing their minds more or less wash out, breaking about evenly between the Republican candidate and the Democratic candidate. In 2016 something very different happened. In 2016, inconsistent responders in the Pew study voted for Trump by a 16-point margin. That is more than double the second largest margin observed in this time series for inconsistent responders (+7 points for George W. Bush in 2000).

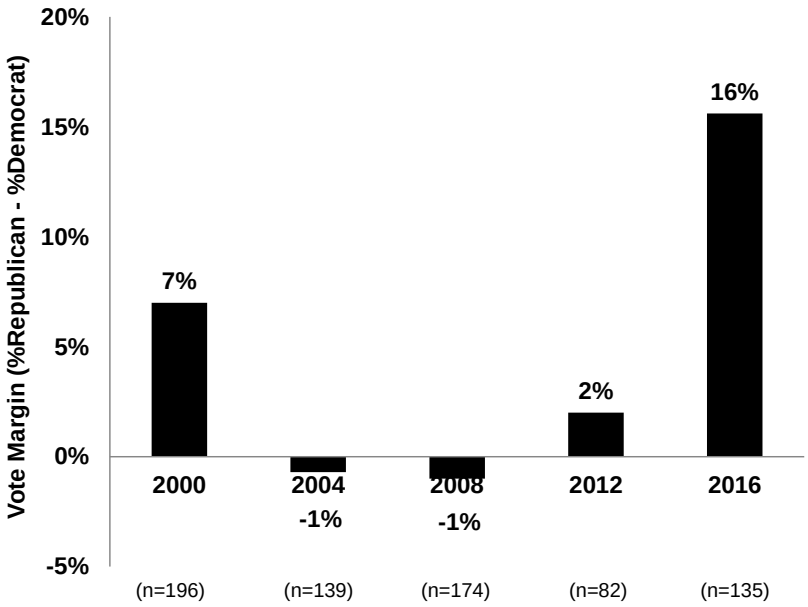


Figure 7. Vote Margin (%Voted for Republican Candidate - %Voted for Democratic Candidate) among Callback Respondents Giving Inconsistent Pre- vs. Post-Election Responses. Note – Data are from Pew Research Center RDD callback studies.

Another way to evaluate this is with the crosstabular data in Table 5. That data shows that 10 percent of all the callback study respondents who ultimately voted for Trump said something different in the pre-election poll. The plurality of inconsistent responders who voted for Trump had described themselves in the pre-election poll as Gary Johnson supporters, about a third had described themselves as undecided or refused to answer, and the remainder had described themselves as supporting some other candidate. Clinton, by contrast, picked up only about half as many late-revealing voters as Trump in this study.

Table 5. Pre-election Poll Responses by the Candidate Ultimately Supported

	Reported Vote			
Pre-election vote preference	Voted for Clinton	Voted for Trump	Voted for other candidate	DK or Refused
Clinton/Lean Clinton	94	1	15	26
Trump/Lean Trump	1	90	4	45
Johnson/Lean Johnson	2	4	41	0
Stein/Lean Stein	1	1	25	3
Other candidate	0	1	11	3
DK-Refused to Lean	2	3	5	23
	100%	100%	100%	100%
Interviews	(587)	(533)	(103)	(31)

Source: Pew Research Center 2016 Election Callback Study. Based on 1,254 completed re-interviews with survey respondents who said that they voted in the general election. Estimates are unweighted.

As discussed above, to describe the inconsistency as "misreporting" would not necessarily be correct because *undecided* or *leaning to Gary Johnson* may have been an accurate answer at the time of the pre-election poll. Regardless, the net effect on an election projection based on such a pre-election poll would be an error of roughly two percentage points in under-estimating support for Trump. Clinton's estimated national popular vote lead based on the responses people in this study gave before the election was 6 percentage points, and her national lead based on those same individuals' post-election responses was 4 points.[11]

In addition, a small percentage of those screened for the post-election callback survey reported not voting (about 8 percent, *n*=104). Clinton led Trump 44 percent to 27 percent among those who reported not voting. Thus, nonvoting hurt Clinton slightly more than it hurt Trump among this small sample.

3.2 Reporting Error and More about the *Shy Trump* Hypothesis

Another widely discussed hypothesis about polling errors in 2016 is the *Shy Trump* effect. The *Shy Trump* hypothesis is a variation on what is generally called the *Shy Conservative* hypothesis in other countries (such as the U.K.). In most election polling misses, the conservative side has been under-estimated more often than the more progressive/liberal side (Jennings and Wlezien 2016). However, historically this has not been the case generally in the United States (see section 2.1). The *Shy Trump/Conservative* hypothesis has its roots in Elizabeth Noelle-Neuman's famous *Spiral of Silence* hypothesis which states that "under the pressure of a hostile opinion climate (national, local, or group level) individuals are reluctant to voice their opinions on morally loaded issues" (Bodor 2012). However, research has generally failed to validate the existence of a spiral of silence, except in some very specific contexts (Bodor 2012).

If Trump supporters refrained from revealing their vote more so than supporters of other candidates, they may have tended a) not to reveal any preference or b) reveal a preference considered more socially acceptable. This reaction should be more present in interviewer-administered than self-administered surveys because the former involves revealing preferences to another person. Therefore, if a *Shy Trump* effect did in fact contribute to polling errors there are several patterns that we would expect to observe.

- The estimates of Trump's support should be lower in live-interviewer telephone polls than in self-administered polls (online and IVR).
- There should be a relationship between estimates of support for Trump in the polls and the proportion of non-disclosers (comprising undecideds and refusals). No such relationship should exist for the other candidates.

3.2.1 Comparing Trump Support in Interviewer- versus Self-Administered Polls

We examined polls to see whether interviewer-administered polls elicited lower estimates of Trump support than self-administered polls. For this analysis, we use the dataset of 208 battleground and 39 national polls conducted during the final 13 days of the campaign (section 2.4). The analysis showed that interviewer administered polls did not under-estimate Trump's

[11] The estimated 6-point Clinton lead based on pre-election data reflects responses from all registered voters weighted with Pew Research Center's standard protocol for general population surveys.

support more than self-administered IVR and online surveys, a finding that is inconsistent with the Shy Trump theory. Battleground state polls with live interviewers were actually among the least likely to under-estimate Trump's support (average signed error of 1.6 points), higher than IVR surveys (0.9) but lower than polls using IVR + Internet administration (2.3) or internet-only administration (3.2). At the national level, live interviewer polls exhibited little systematic error under-estimating Trump's vote margin (0.4), while under-estimation was slightly higher for Internet modes (1.1) and IVR/Internet surveys over-estimated Trump's support slightly (-0.7 signed error). This pattern is mirrored by results from the regression analysis of mode and other factors on absolute error, which found that only one self-administered mode (IVR) was associated with lower errors than live phone interviewers.

If the *Shy Trump* effect was real, however, there is no reason to expect that it would have been confined to polls conducted very late in the campaign. Presumably, any hesitation about disclosing support for Trump would have been just as pronounced (if not more so) in September and early October. Thus, we also tested for this mode of administration difference using published polls conducted from September 1st to Election Day. With this larger set of polling data, we were also better able to apply more sophisticated statistical tests.

Figure 8 shows the national trend in voting intentions for Trump, by mode, using a local regression estimation. It illustrates that estimates produced by live telephone polls were similar to those produced by self-administered Web polls. The mode that stands out somewhat is IVR + Internet, which tended to show Trump garnering about 50 percent of the major party vote. Estimates of Trump support from the two other modes tended to be about 2.5 percentage points lower.

However, these aggregate effects may be due to other features of the polls than just mode of administration, hence the necessity for refined statistical testing. To better isolate an effect from mode, we conducted a regression analysis that controls for length of field period, tracking poll versus non-tracking poll, likely voter (LV) versus registered voter (RV) estimate and change over time (Appendix A.E). The results were highly consistent with the analysis just using polls from the final 13 days. Self-administered online polls and interviewer-administered phone polls both recorded *lower* levels of support for Trump than IVR polls.

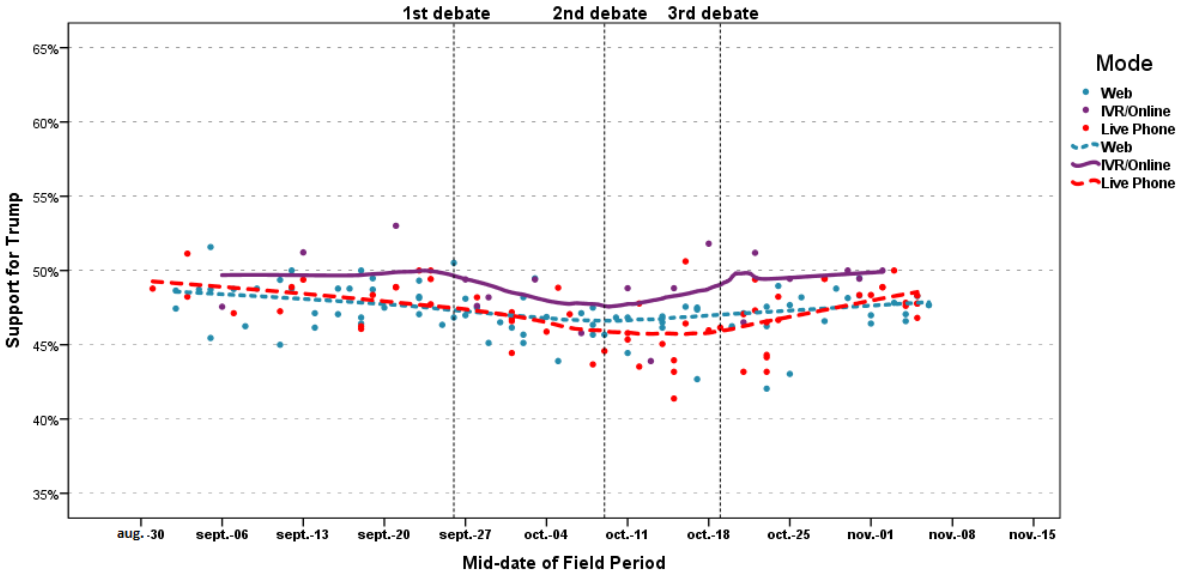


Figure 8. Support for Trump (On the Sum of the Two Major Party Candidates) by Mode.
Note – Each point represents a poll estimate positioned at the midpoint of the field period. Lines represent Loess estimates of change over time using Epanechnikov .65 estimation. © C. Durand, 2016.

The finding that live telephone surveys did not consistently under-estimate Trump's support more than self-administered online polls is informative, though not conclusive, evidence against the *Shy Trump* hypothesis. Live telephone polls and self-administered polls differ by too many important factors (e.g., sample source, weighting) for this type of analysis to cleanly isolate the effect from interviewer presence, even when using statistical modeling. That said, the results are inconsistent with expectations of the *Shy Trump* theory.

3.2.2 Comparing Nondisclosure in Self- versus Interviewer-Administered Polls

One possibility is that Trump supporters were more likely than other respondents either to report being undecided or to refuse to reveal their preference. In that case, we would expect to observe a relationship between the proportion of nondisclosers and the proportion of Trump supporters in the polls and no such relationship for Clinton. However, the proportion of nondisclosers is related to the methodological characteristics of the polls. The average rate of nondisclosure was highest for online polls (8.5 percent) and lower for IVR + Internet (5.6 percent) and live phone (4.3 percent).[12]

Appendix Table A.9 shows that the proportion of non-disclosers in polls is not related to the proportion of support for Trump, all else being equal. However, if we consider the estimates for all the candidates, we see that polls that had larger shares of nondisclosers showed more support for both Trump and Clinton and less support for third party candidates. The main takeaway is that there is no evidence that higher rates of undecided or refusals to answer (that is,

[12] This difference is thought to be basically an artifact of the difficulty of treating "leaners" the same way on the phone and on the web (Cohn 2016a). In phone polls, if a respondent refuses to say they plan to vote for one of the candidates mentioned, then they are typically asked a follow up question asking which candidate they are *leaning* towards. This follow-up format tends to yield relatively low levels of nondisclosure. In online polls, by contrast, such a follow up format is less common.

nondisclosure) is associated with level of Trump support, thus failing to yield evidence supporting the Shy Trump hypothesis.

3.2.3 Experiments Testing the Effect of Mode on Support for Trump

In 2016, one polling organization, Morning Consult, conducted two experiments designed to isolate the effect of self- versus interviewer-administration (Dropp 2016) on support for Trump. While the first experiment was conducted in the run up to the primaries and the second during the general election, they used the same basic design. A group of likely voters was recruited from an online opt-in sample source and asked a set of background questions. They were then randomly assigned to complete the remainder of the interview by either proceeding with an online survey or dialing into a call center and answering questions from a live interviewer. The general election edition of the experiment yielded a mode difference in the expected direction (Clinton +5 points in the live phone condition versus +3 points in the web condition), but the result was not statistically significant. Dropp did report a statistically significant mode effect in the expected direction (more Trump support in the online condition than the live telephone condition) among well-educated and higher-income voters.

More recently, Pew Research Center (2017) conducted an experiment that randomized mode of interview on the Center's American Trends Panel, which is recruited from national landline and cell phone RDD surveys. Half of the panelists were assigned to take the survey online and the other half via a live phone interview. That study, conducted February 28-March 12, 2017, found little evidence that poll participants were censoring support for Trump when speaking to an interviewer. There was no significant difference by mode of interview on any of four questions asking directly about Trump (e.g., presidential job approval, personal favorability). Questions asking about major policy priorities of the Trump administration also showed no mode effect, except on treatment of undocumented immigrants, which showed 8 percentage points more support for the conservative position online relative to on the phone.

As with the other analyses presented in this report, the experiments have their limitations. While Dropp's results may generalize to other polls conducted with online opt-in samples, it is not clear how well they generalize to other polls with samples from a voter file or RDD samples. It is also not clear whether differential nonresponse to the latter part of the interview posed a threat to the mode comparison. It seems likely that break off was higher in the phone condition than the web condition, but how well that could have been corrected through statistical modeling is not clear. For its part, the Pew study speaks more directly to polls conducted since Trump took office, than it does to 2016 pre-election polls. As noted in the report, the timing of that study (conducted more than one month after Trump took office) and the fact that it was not focused on presidential vote means that it only indirectly speaks to the possibility of a *Shy Trump* phenomenon in 2016.

3.2.4 Using a Question about Neighbors' Vote to Adjust for *Shy Trump* Responding

As discussed above, polls generally under-estimated Trump's support in Pennsylvania and Michigan – but there was one exception. Trafalgar Group, a Republican-affiliated IVR firm, was the only pollster to correctly project Trump victories in both states. In fact, in each of the six battleground states they polled, they over-estimated support for Trump. In states like Michigan, Pennsylvania and North Carolina, Trafalgar's pro-Trump tilt yielded impressive results. But in Colorado and Florida, the over-estimation of Trump support led to larger absolute errors (3.9 and

2.8 points, respectively), albeit with numbers that projected the correct winner. While Trafalgar did forecast Trump wins in both Michigan and Pennsylvania, they were not necessarily the most accurate pollster or even the most accurate IVR poll in 2016.

Two distinctive design decisions seem to explain why Trafalgar's results were consistently more favorable to Trump. They took a novel approach for producing final vote preference estimates. According to their methods report, "the final published ballot test is a combination of survey respondents to both a standard ballot test *and a ballot test gauging where respondent's neighbors stand*. This addresses the underlying bias of traditional polling, wherein respondents are not wholly truthful about their position regarding highly controversial candidates." (emphasis added) The general idea is that if people will not admit they personally would vote for Trump, they would admit that their neighbors would. As Stinson (2016) reported, the other distinctive feature of Trafalgar's polling was that they selected their samples from voter files using a more-inclusive-than-normal approach that included registered voters who had not voted for years. Some had not voted since 2006. According to Trafalgar CEO Robert Cahaly, other pollsters tend not to sample such records.

It is not clear what the relative contributions of these two factors were on the overall performance of the poll. Also, while the Trafalgar methods statement asserts that the incorporation of the neighbor vote intention question is effective because it corrected for *Shy Trump*-type responding (and that may have been the case), it also seems possible that in states like Michigan and Pennsylvania it was correcting some other error (e.g., over-representation of Democratic-leaning college graduates). The methods report suggests that Trafalgar, like a number of other IVR pollsters, did not measure respondent education, so this may remain something of a mystery. Regardless it is informative to observe that these two unusual methodological levers were pulled and they had the effect of overcoming the general pro-Clinton error that seemed to plague most pollsters to varying degrees in 2016. On its face, the practice of using a more inclusive voter file sample that brings in dormant voters seems like something others may want to evaluate. The other idea of integrating reports about neighbors' vote choice with self-reported vote choice also warrants experimentation in a broad array of contests so as to better understand the properties of that measurement approach.

3.2.5 Did Trump Out-Perform Polls More Than Republican Senate Candidates?

A different way to test whether polling errors were attributable, at least in part, to misreporting is to compare Trump's performance in state-level polls to the performance of Republican candidates for Senate in those same polls. Presumably, respondents who may have felt pressure to censor their support for Trump did not feel similar pressure to censor support for the Republican Senate candidate. If such differential censoring did occur, then we would expect to see – at the individual poll level – that Trump outperformed his poll number by a larger margin than the Republican Senate candidate did. Such a result, while not definitive, would suggest that part of the error in the presidential race estimates was attributable to misreporting.

To examine this, we used battleground state polls conducted entirely within the final two weeks of the election. To be included, each poll needed to measure both Senate and presidential vote preference. There were 34 Senate contests in 2016, eight of which were held in states where the presidential vote margin was less than five percentage points (AZ, CO, FL, NH, NV, NC, PA,

WI). We examined the final polls for these eight states and, for each state, used only the last poll conducted by each firm. This yielded an analytic dataset with 66 polls, 24 of which were conducted entirely with live telephone interviewing.

Here we defined "over-performance" as the signed difference between the final vote margin and the poll margin, where the margin is the Republican vote minus the Democratic vote.[13] Table 6 provides an illustration of how these computations were done using one state, Wisconsin. The final Marquette Law School Poll had the Senate margin at -1 (44% for Johnson, the Republican and 45% for Feingold, the Democrat) and the presidential margin at -6 (38% for Trump and 44% for Clinton). The actual election in Wisconsin went +3.4 for Johnson and +0.7% for Trump. In this analysis, Johnson over-performed the Marquette poll by 3.4 - (-1) = +4.4 points, and Trump over-performed by 0.7 – (-6) = +6.7 points. Comparatively speaking, Trump over-performed the poll by 6.7 – 4.4 = 2.3 points more than the Republican Senate candidate did.[14] This difference in differences, the dependent variable in the analysis, is shown in the far right column of Table 6.

Table 6. Illustration of How Over-performance was Measured Using Wisconsin as an Example

	Senate Race		Presidential Race		Over-performance (Vote Margin - Poll Margin)		
Source	Rep. Candidate	Dem. Candidate	Rep. Candidate	Dem. Candidate	Senate (Johnson)	President (Trump)	*Difference (Pres. error - Sen. error)*
Vote Outcome	50.2%	46.8%	47.2%	46.5%			
Loras College	45%	47%	38%	44%	5.4%	6.7%	*1.4%*
Marquette Univ.	44%	45%	40%	46%	4.4%	6.7%	*2.3%*
Emerson College	44%	49%	42%	48%	8.4%	6.7%	*-1.7%*
PPP	44%	49%	41%	48%	8.4%	7.7%	*-0.7%*
SurveyMonkey	49%	48%	42%	44%	2.4%	2.7%	*0.3%*

Most Senate races featured just one or two late live telephone polls. Rather than attempting this analysis separately at the state level (where data are too sparse), we use a combined dataset with results for the 66 polls from eight states. A number of findings in the summary statistics (Table 7) merit discussion. The central question is whether Trump tended to out-perform his poll numbers more than a Republican Senate candidate in the same poll, particularly for live telephone polls. As shown in the first row of Table 7, we find no support for that idea. In the 24 live telephone polls analyzed, Trump beat his poll estimate by 1.4 percentage point on average, and the Republican Senate candidate beat his or her poll estimate by a nearly identical 1.3 percentage points on average. An independent, very similar analysis by Harry Enten (2016) reached the same general conclusion.

[13] We also considered defining "over-performance" with respect to the candidate's estimated share of the total vote, as opposed to using the Republican-Democrat margin as described in the text. We found, however, that vote share was not a suitable framework. Due to the fact that polls feature undecided voters and tended to over-estimate support for third party candidates, both Donald Trump *and* Hillary Clinton generally "over-performed" relative to their estimated vote share in polls.

[14] The level of imprecision or, more specifically, the standard error of these estimates is worth considering though not addressed in this report. Given that this analysis yielded basically a null result the standard errors are, for practical purposes, a moot point.

Table 7. Trump's Over-performance of Polls Relative to Republican Senate Candidates in Battleground States

	Average Over-performance (Vote Margin - Poll Margin)			
Type of Poll	Senate Rep. Candidate	President Rep. Candidate	*Ave. Difference (Pres. error - Sen. error)*	Polls
Live phone	1.3%	1.4%	0.0%	24
Online	4.5%	3.2%	-1.3%	17
IVR, IVR+Online	2.7%	1.8%	-0.9%	22
Other	7.7%	3.9%	-3.8%	3

Also, as election observers will recall, not only did Trump out-perform poll estimates, so did most Republican candidates in competitive Senate races. This pattern is evidenced by the fact that all of the values in the first column are positive. This finding is suggestive of systematic under-estimation not just of support for Trump but of Republican candidates more generally. Indeed, Republican candidates for the U.S House of Representatives also tended to outperform their poll numbers. Nationally, the actual congressional vote was +1.1 for Republicans, whereas the final polling average from RealClearPolitics was estimated at +0.6 for Democrats. The fact that polls tended to under-estimate support for Republican candidates writ large in 2016 – not just support for Trump – undermines the notion that polling errors were caused by socially desirable reporting.

3.2.6 Effects of Interviewer Characteristics on Presidential Vote Preference

Another indirect test for socially desirable reporting is to look at whether responses to the vote preference question varied by potentially discernable interviewer characteristics, such as gender and race. For example, if poll respondents interviewed by white males were significantly more likely to report intending to vote for Trump than those with female and/or non-white interviewers, that would suggest misreporting was a problem. It is possible that some respondents who knew they were Trump voters were reluctant to say so, even to white male interviewers, so this is an imperfect test.

Two microdatasets made available to the committee contained variables for interviewer race and gender, the ABC News/Washington Post poll and Pew Research Center's October poll. While simple bivariate analysis seems to suggest some effect from interviewer characteristics (the margin was +2 Clinton among interviews completed by non-white interviews versus -1 Clinton among interviews by white interviewers in the ABC News/Washington Post poll), no meaningful effects were detected. Because interviewers are not randomly assigned to respondents, statistical models are required to estimate the effects of interviewer race and sex on respondent vote preferences. In multivariate modeling, if one controls for basic respondent demographics (gender, race/ethnicity, education) any effect from interviewer race or interviewer gender disappeared (i.e., became nonsignificant). The lack of any evidence for an effect of interviewer race or gender on how respondents answered the presidential vote question is not conclusive evidence against the *Shy Trump* hypothesis. However, the result is inconsistent with expectations of the *Shy Trump* theory, and suggests other factors than socially desirable reporting were responsible for the bulk of error in general election polls.

3.3 Nonresponse Bias

One alarming possibility raised by the direction of polling errors was that, broadly speaking, some segment of Trump's support base was not participating in polls. Participation in polls is quite low across the ideological spectrum and has been for some time (Pew Research Center 2012), and even the most rigorous polls in 2016 had single digit response rates. So it is not the case that *most* Clinton supporters were taking polls while *most* Trump supporters were not; rather the pattern would have been much more subtle. Differential nonresponse, if it was a real problem, would have manifested as Trump supporters being somewhat less willing to participate in surveys, on average, than Clinton supporters.

3.3.1 Were National Polls Accurate Because Regional Errors Offset?

While national polls clearly performed better than state-level polls on average, at least one set of commentators suggested that the strong performance of national polls was a mirage. Cohn, Katz and Quealy (2016) observed that Trump out-performed his poll numbers in states with a large number of white voters without a college degree and under-performed his polls numbers in large, liberal states with sizable Hispanic populations. Overall, they noted, "the two types of misses nearly canceled out in national polls." If true, then the conclusion reached here and elsewhere that the national polls were generally accurate while many state polls were not would be discredited. If the low error on national polls was simply a fortuitous outcome of two large errors canceling, then it would be more accurate to conclude that neither state-level nor national-level polls did a good job of capturing the voting electorate in the 2016 general election.

To test whether national polls appeared to perform well simply because large errors canceled, we used final microdatasets from three RDD polls (ABC News/Washington Post, CNN/ORC, Pew Research Center) and one online opt-in poll (SurveyMonkey). If the assertion were true, we would expect to find that these national polls noticeably under-estimated Trump's support in key working class white states (PA, MI, MN, WI, OH) while simultaneously over-estimating Trump's support in liberal, relatively Hispanic states (CA, NY, NV, IL, WA). The results are presented in Table 8. It must be noted that of these four surveys, only one (SurveyMonkey) was designed for state-level inference and released state-level vote estimates. In addition, two of the polls (CNN/ORC and Pew Research Center) were conducted at least two weeks before Election Day. While it is, therefore, unrealistic to expect the subnational estimates for all of these polls to align perfectly with the vote, we felt it was reasonable to check the data for the general pattern in question.

Table 8. Trump Margin by Region

	Liberal and Relatively Hispanic States	Other States	Competitive, White Working Class States
Actual vote margin (T-C)	-23%	5%	2%
Pew Research Center			
Poll margin (T-C)	-30%	1%	2%
Difference from vote	**-7%**	**-4%**	**0%**
N	489	1,284	347

CNN/ORC			
Poll margin (T-C)	-23%	5%	-13%
Difference from vote	**0%**	**0%**	**-15%**
N	181	475	123
ABC News/Washington Post			
Poll margin (T-C)	-21%	4%	-2%
Difference from vote	**2%**	**-1%**	**-4%**
N	761	1,957	493
SurveyMonkey			
Poll margin (T-C)	-23%	0%	-1%
Difference from vote	**0%**	**-5%**	**-3%**
N	10,150	51,648	12,388

Sources: ABC News/Washington Post RDD tracking poll interviews from November 1-7, 2016, Pew Research Center RDD survey fielded October 20-25, 2016, CNN/ORC RDD survey fielded October 20-23, 2016, SurveyMonkey interviews fielded November 1-7, 2016. Note: Some differences do not sum due to rounding. States coded as "Liberal and Relatively Hispanic" were CA, NY, NV, IL and WA. States coded as "Competitive White Working Class" were PA, MI, MN, OH and WI.

Overall, the data are not consistent with the claim that the relatively accurate results in national polls in 2016 resulted from two large errors canceling each other out (over-statement of Trump support in liberal, heavily Hispanic states and understatement in working class white states). For any given poll, that narrative only gets about half the story correct. The CNN/ORC poll accurately estimated Trump support in predominantly liberal and Hispanic states; all the projection error was in Upper Midwest. For the Pew Research Center poll, the opposite was true; the projection error was predominantly in the liberal, Hispanic states. For the ABC/Washington Post and SurveyMonkey polls, the estimates of Trump support were too low in the Upper Midwest (relative to the outcome), but Trump's margin of defeat in liberal, Hispanic states was smaller or the same in those polls than in the vote.

In fairness, if one looked at how the state-level polls performed and assumed that national polls are basically conducted the same way, then this theory of canceling errors seemed very plausible. It overlooks one key point, however. State-level polls and the national polls are not conducted the same way. As discussed in section 2.4, live telephone interviewing represents a much larger share of national polls (36%) than state-level polls (18%). While an IVR or online opt-in poll may cost in the neighborhood of $5,000 to $15,000, live telephone polls with professional interviewers cost closer to $100,000 (Cassino 2016). This means that the resources going into a typical state poll can be dwarfed by those that go into a national poll. In addition, national pollsters are nearly twice as likely to adjust for education as state-level pollsters.

3.3.2 Did Polls Under-represent Staunchly Pro-Trump Areas?

One way to test for differential partisan nonresponse is to leverage information about which parts of the country were staunchly pro-Trump and how many people live in those areas versus the rest of the country. If polls systematically failed to interview people in staunchly pro-Trump areas,

we would expect to find residents of such counties under-represented in polls. For example, if the Census shows that 13% of Americans live in staunchly pro-Trump areas, but polls estimate that only 9% of Americans live in those same areas – that would be evidence that polls were, indeed, systematically missing Trump supporters. Somewhat surprisingly (given the polling errors), we found no evidence to that effect. The results are presented in Table 9.

Since there was no obvious, definitive way to define a "staunchly pro-Trump" area, we tested three definitions. The definition used in the first row of the table identifies counties in which Trump won by at least a 40-point margin. The definition used in the second row identifies counties in which Trump won by at least a 60-point margin. Finally, third row simply identifies rural counties, defined as those with a population density of fewer than 50 people per square mile. The rural definition was motivated by the fact that Trump, like most Republican presidential candidates, generally had much stronger support in rural areas than metropolitan areas. Census estimates for the share of the population living in areas identified using each of these three definitions come from the 2015 Census population estimates. Poll estimates come from two microdatasets that contained the requisite county-level information – the mid-October CNN/ORC poll (n=1,017) and a cumulative dataset with all 15,812 telephone interviews Pew Research Center conducted in 2016 political polling.[15]

Table 9. Estimates of the Share of U.S. Adults Living in Staunchly Pro-Trump Counties

		Share of the U.S. Population Living in Those Areas				
			CNN/ORC poll		Pew Research poll	
Three definitions of staunchly pro-Trump areas	Number of counties	Census Benchmark	Weighted estimate	Unweighted estimate	Weighted estimate	Unweighted estimate
Counties Trump won by 40+ points	1,486	13%	16%	16%	13%	13%
Counties Trump won by 60+ points	524	3%	4%	4%	3%	3%
Rural counties (< 50 people/mi^2)	1,657	9%	12%	12%	9%	10%

[15] As stated in the footnote of Table 9, the Census figures are based on all ages and the CNN/ORC and Pew Research Center figures are based on all adults age 18 or older. We investigated whether that discrepancy confounded the comparison in a noticeable way and concluded it did not. While it seemed possible that rural and other staunchly pro-Trump areas skew slightly older than other parts of the country, we did not see empirical evidence of that. For example, the predominantly rural and overwhelmingly pro-Trump states of Oklahoma and Wyoming represented equal shares of the entire U.S. population (1.2% and 0.2%, respectively) and the U.S. adult population (also 1.2% and 0.2%, respectively). Consequently, we concluded that this small discrepancy has no meaningful impact on the results or conclusions in this analysis.

Note - The Census figures are based on people of all ages. Source: Census figures are 2015 population estimates. CNN/ORC estimates based on 1,017 interviews conducted October 20-23, 2016. Pew data are based on a cumulated file with all 15,812 interviews conducted in routine dual frame RDD surveys in 2016. The CNN/ORC and Pew figures are based on people age 18 or older.

If the polls were systematically missing people in staunchly pro-Trump areas, then the figures in the unweighted estimate columns would be noticeably lower than the Census benchmarks in the second column. If such a pattern was not fixed by the weighting, then the estimates in the weighted estimate columns would also be noticeably lower than the Census benchmarks. Neither of those patterns is present in the data. If anything, people living in the most pro-Trump parts of the country are slightly over-represented.

These findings do not rule out the possibility that differential nonresponse was a factor in polling errors in 2016. For example, it is possible the people interviewed in these pro-Trump areas were not representative with respect to their vote choice. It is also important to note that this analysis, based on telephone RDD polling data may not generalize to online opt-in polls or IVR polls. Even with these caveats, it is informative that this particular test, which we expected might detect under-representation of pro-Trump areas, does not show evidence of bias.

3.4 Weighting

One hypothesis about 2016 polling errors is that pollsters did not interview enough white voters without a college degree (Silver 2016b). Indeed, many pollsters are likely to acknowledge that contemporary polls *almost never* interview enough voters without a college degree. Numerous studies have shown that adults with less formal education tend to be under-represented in surveys on an unweighted basis (Battaglia, Frankel and Link 2008; Chang and Krosnick 2009; Link et al. 2008; Pew Research Center 2012). A seasoned pollster would be quick to emphasize, however, that this well-established education skew need not bias their estimates. Many pollsters adjust their samples to population benchmarks for education in order to address this very issue. As long as the pollster accounts for the under-representation of less educated adults in their weighting, then this issue would not lead to bias, so long as the less educated adults they did interview were representative of the ones they did not interview.

In the weeks following the 2016 general election, however, one intriguing fact started to emerge: not all pollsters, particularly those polling at the state level, adjusted their weighting for education. Why would that have undermined polls in 2016 but not previous elections? The answer is that in 2016 the presidential vote was strongly and fairly linearly related to education; the more formal education a voter had, the more likely they were to vote for Clinton (see the right-hand panel in Figure 9). Historically, that has not been the case. In most modern U.S. elections, presidential vote (defined here as support for the Democratic candidate) exhibited a U-shaped or “curvilinear” pattern with respect to education. For example, as shown in the left-hand panel of Figure 9, in 2012 both the least educated and most educated voters broke heavily for Barack Obama, while those in the middle (with some college or a bachelor’s degree) split roughly evenly for Mitt Romney and Barack Obama.

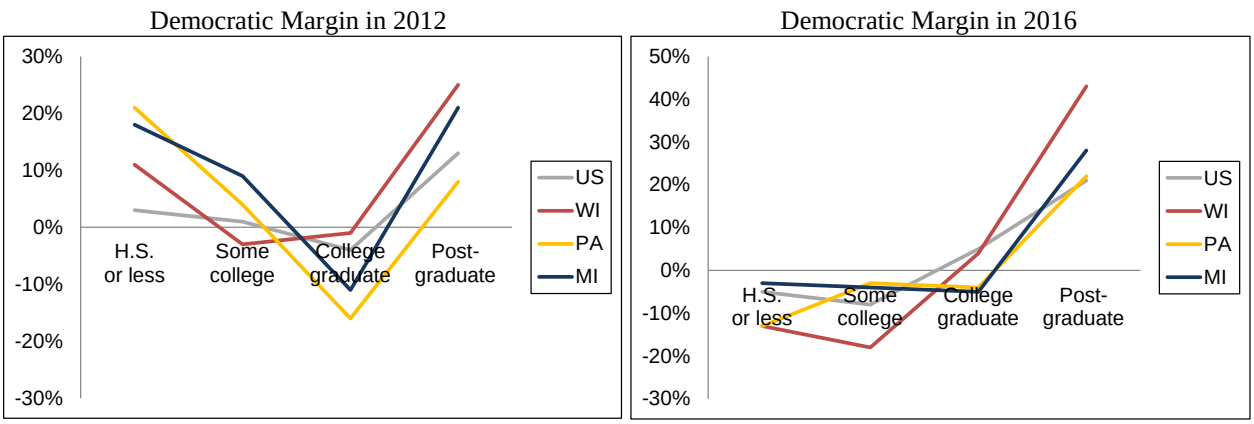


Figure 9. Democratic Presidential Vote Margin in 2012 and 2016 by Voter Education Level and Geography. Source: NEP national Exit Poll 2012, 2016

To understand why pollsters could perhaps get by without weighting on education in an "U-shape" election like 2012 but not a linear election like 2016, consider the post-graduate results. In a U-shaped election, the post-graduate voters who are likely to be over-represented in polls that are not adjusted for education vote in much the same way as the low-education voters that such polls under-represent. By contrast, in 2016, that completely fell apart. In 2016, highly educated voters were terrible proxies for the voters at the lowest education level. At least that was the case nationally and in the pivotal states in the Upper Midwest.

Following the election, two different state-level pollsters acknowledged that they had not adjusted for education and conducted their own post-hoc analysis to examine what difference that would have made in their estimates. Both pollsters found that adjusting for education would have meaningfully improved their poll's accuracy by reducing over-statement of Clinton support.

The final University of New Hampshire (UNH) poll had Clinton leading in the Granite State by 11 points. She ultimately won by a razor thin 0.4-point margin. The UNH poll director, Andrew Smith (2016), reported that the released poll adjusted for age, gender and region but not education – a protocol that had served the Granite State Poll just fine for numerous election cycles. According to Smith (in email correspondence), "We have not weighted by level of education in our election polling in the past and we have consistently been the most accurate poll in NH (it hasn't made any difference and I prefer to use as few weights as possible), but we think it was a major factor this year. When we include a weight for level of education, our predictions match the final number." Indeed, as shown in Figure 10, had the UNH poll adjusted for education in 2016, that single modification would have removed essentially all of the error. The education-adjusted estimates showed a tied race.

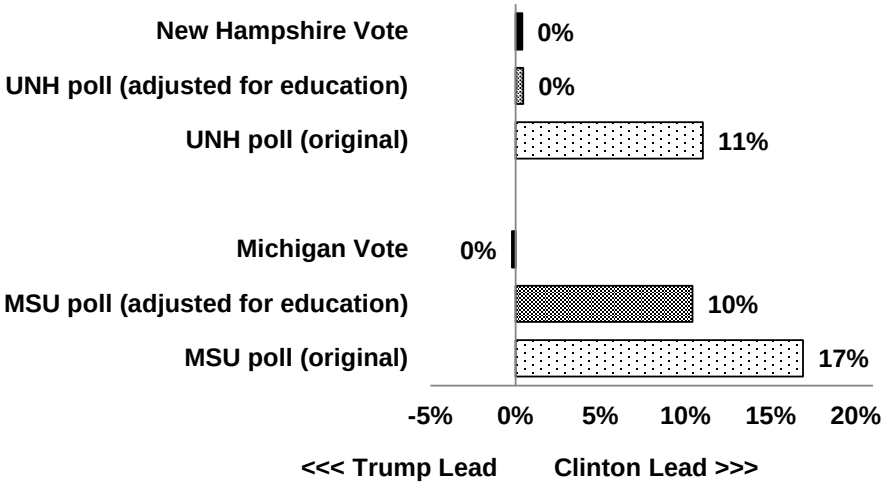


Figure 10. Poll Estimates with and Without Weighting Adjustment for Education, Relative to 2016 Presidential Vote Outcome. Source – University of New Hampshire poll conducted November 3-6, 2016 with 707 likely voters. Michigan State University poll conducted September 1 – October 30, 2016 with 743 likely voters.

The story is similar, though less dramatic, for Michigan State University's (MSU) State of the State Poll. That poll, which like the UNH poll was conducted via live phone with a dual frame RDD sample, showed Clinton leading Trump in Michigan by 17 points.[16] She ultimately lost that contest by another slim margin (0.2 points). The MSU poll did not adjust for education, but if it had, Clinton's estimated lead would have been 10 points, instead of 17. One other noteworthy feature of the MSU poll is that, unlike the UNH poll, it was fielded relatively early, with most interviews completed before mid-October. This means that the MSU poll largely missed what appears to be a significant, late shift in support to Trump. As discussed in Section 3.1, the national exit poll indicates that about 13 percent of Michigan voters made their presidential vote choice in the final week of the campaign, and that group went for Trump by about an 11-point margin.

It was not just RDD pollsters who, in hindsight, would like to have handled education differently in their weighting. SurveyMonkey's Head of Election Polling (and report co-author), Mark Blumenthal (2016), reported their online opt-in poll weighting did adjust for education but used three categories that were quite broad (high school or less, some college, and college graduate). According to Blumenthal, "If we had separated out those with advanced degrees from those with undergraduate degrees in our education weighting parameters, we would have reduced Clinton's margin in our final week's tracking poll by 0.5 percentage points to +5.5 (47.0% Clinton to 41.5% Trump)."

Despite this, it is not clear that adjusting to a more detailed education variable would have universally improved polls in 2016. Analysis of the effect from weighting by five education categories rather than three categories in four national polls (Appendix A.H) yielded an average

[16] An early release of the MSU poll reported a 20-point Clinton lead (http://msutoday.msu.edu/_/pdf/assets/2016/state-of-state-survey.pdf). The corresponding microdataset provided to the committee, presumably reflecting the final release, gives a 17-point Clinton lead as shown in Figure 10.

change of less than 0.4 percentage points in the vote estimates and no systematic improvement. In sum, the difference between weighting to education or not (as with UNH and MSU) is much more dramatic than the difference between weighting to more or less detailed education categories (as with SurveyMonkey).

Given the evidence that not adjusting on education led to an unintended pro-Clinton bias in several polls – might this explain more generally the polling errors observed in the general election? For opt-in online polls the answer appears to be no. Adjusting on education did not seem to differentiate opt-in online polls in terms of accuracy. This was something of a surprise given how large the effect of weighting on education seemed to be in the RDD polls discussed above.

Isolating the effect of weighting on education was difficult because the online opt-in polls with microdata available to the committee all adjusted for education. If weighting on education was an important factor in the accuracy of online opt-in polls, we would expect to see evidence that polls making that adjustment produced more accurate results than those not making that adjustment. There were five online opt-in pollsters who released late estimates in each of seven contests (the presidential race in AZ, FL, GA, NC, OH, PA, and the national popular vote). Of these pollsters, three weighted on education and two did not. As shown in Figure 11, there is no indication that weighting on education was associated with accuracy (i.e., lower absolute error). The two online polls that did not adjust on education performed about the same as the polls that did adjust on education. Admittedly, this is a crude test as these five online opt-in polls differ on other factors (e.g., question wording, sample design) besides adjusting on education. Still, if weighting on education was an important reason why some online polls were off, it seems reasonable to think that might manifest in this comparison.[17]

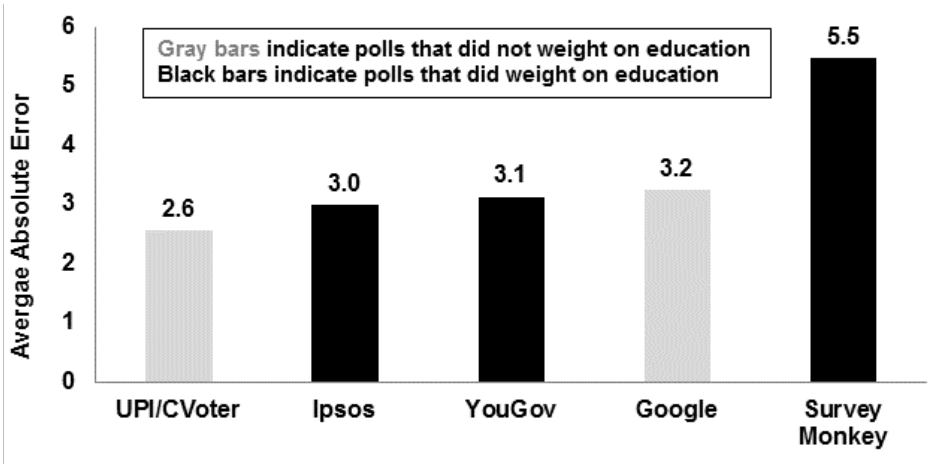


Figure 11. Average Absolute Error on 2016 Presidential margin in Seven Shared Contests, by Five Online Opt-in Polling Organizations. The seven contests for which each of these organizations released estimates in the final two weeks were AZ, FL, GA, NC, OH, PA, and the U.S.

[17] Some online opt-in polls adjust for education at the sampling stage through quotas or sample matching, which can mitigate the need to adjust for education in weighting. To the best of our knowledge, neither UPI/CVoter nor Google Consumer Surveys (shown in Figure 11) adjusted for education at the sampling stage.

While it is not clear that adjusting for education matter greatly for online polls, it may have been important for the accuracy of IVR polls just as it seems to have been for RDD. About 40 percent of state-level polls were conducted via IVR in part or in whole, but to the best of our knowledge none of those microdatasets are publicly available, and private requests for microdata on behalf of the committee were turned down or ignored. Unfortunately, many pollsters, particularly those not participating in AAPOR's Transparency Initiative, do not report the variables they use in weighting. Some but not all such pollsters disclosed this information upon request.

We attempted a poll-level analysis to examine the effect of adjusting on education or not. That effort was basically unsuccessful due to the sparseness of the data. The only finding from that effort that we think is reliable is an assessment of how many pre-election polls appear to have adjusted for education in their weighting. This effort required manual investigation and coding of each poll, and so it was performed on a subset of the state contests (FL, OH, MI, NC, PA, WI) in addition to national polls. We considered only polls conducted in the final two weeks and only each pollster's final poll (to avoid double counting pollsters who fielded more than one poll in the final weeks). This yielded an analytic dataset with 102 polls. Despite individual pollster outreach efforts, we were unable to determine whether 17 of these polls had adjusted on education. Most of these polls (15 of the 17) featured at least some IVR and at least some voter file sample. Virtually all polls of this type that *did* disclose their weighting variables *did not* adjust on education. We, therefore, felt reasonably comfortable assuming that polls with missing weighting information did not adjust for education in this analysis. We allow that this imputation rule may be incorrect for a handful of polls, but based on the data that are present we think it is highly unlikely that the imputation rule is wrong for a meaningful number of polls in this analysis.

Table 10. Share of Pollsters That Adjusted on Education in Weighting

Type of Poll	Share of polls that weighted for education	Number of final polls
Michigan polls	18%	11
Wisconsin polls	27%	11
North Carolina polls	29%	14
Florida polls	31%	16
Pennsylvania polls	33%	18
Ohio polls	36%	11
National polls	52%	21

Note - Figures reflect only polls fielded in the final two weeks and only a given pollster's final poll. The requisite weighting information was missing for 23 polls, which were all imputed as not weighting on education, based on information among similar polls that did disclose their weighting variables.

Table 10 shows that most state-level polls do not adjust for education in weighting, whereas about half of national polls do. In Michigan, only one-in-five polls adjusted for education, while in Ohio just over one third (36 percent) did so. The polls in other decisive states fell somewhere in between. Our impression from reviewing the polls is that the main contributor to this pattern is that in 2016 the most common (or "modal") state-level poll was an IVR poll that drew its sample

from a voter file and may or may not have fielded a supplemental opt-in online sample. One advantage of sampling from a voter file is that the pollster has useful information about the poll respondents and nonrespondents. This information, which is frequently used in weighting adjustments, includes voter age, gender, geography, party registration, past voting history and, for some states, race. Some polls also adjust their weights with modeled data for the likelihood of voting. One variable that is not on the voter file and is absent from weighting protocols of polls sampling off the voter file is education.

Pollsters who sample from the voter file could adjust for education using some other source, such as the Current Population Survey, but most of them choose not to. Indeed a number of IVR pollsters who sampled from voter files by all appearances did not even measure respondent education much less adjust for it in weighting. Table 11 shows that in Michigan, Pennsylvania, and Wisconsin only about half of the IVR pollsters were measuring respondent education, based on the toplines in their press releases (which show other demographics like gender and race).

Given the large error associated with education imbalance in some RDD surveys, it seems quite possible that this presented a problem for IVR pollsters as well. While no definitive data are available for the demographic profile of the voting electorate in 2016, the NEP exit poll could be used as a stand-in. Cohn (2016b) reported that the NEP over-represents both college educated voters and non-whites. We might, therefore, use the college graduate rate from the NEP *as an upper bound* for the probable rate of voters who were college graduates. Comparing those rates to IVR polls shows that IVR polls seem likely to have over-represented college graduates in critical Upper Midwest states. Given that higher education levels were strongly associated with support for Clinton, this seems likely to have contributed to errors in these states.

Table 11. Share of College Graduates in IVR Polls Relative to NEP Exit Poll in Three States

Michigan		Pennsylvania		Wisconsin	
NEP Exit Poll	*43%*	*NEP Exit Poll*	*48%*	*NEP Exit Poll*	*45%*
Gravis	53%	Gravis	57%	Emerson	48%
Emerson College	48%	Emerson College	54%	Mitchell	N/A
Mitchell Research	N/A	Harper	54%	Trafalgar	N/A
Trafalgar Group	N/A	Trafalgar Group	N/A	PPP	N/A
EPIC/MRA	N/A	PPP	N/A		
PPP	N/A				

Note -- Data come from pollster press releases and appear to be weighted. "N/A" indicates that respondent education level does not appear to have been measured in the poll.

3.5 Turnout Assumptions and Likely Voter Modeling

3.5.1 How Voter Turnout Changed from 2012 to 2016

Changing voter turnout was a part of the story of how Trump defeated Clinton in 2016. Turnout appears to have increased more among Republicans and rural voters in key states. At the same time, turnout was stagnant or lower among some core Democratic voters, particularly African

Americans. The question for pollsters is whether the dynamics of voter turnout in 2016 introduced error into polling estimates of the presidential contest. The assumptions that pollsters make about turnout vary widely. Many do not routinely disclose the demographic composition of their samples and, as of this writing, critical benchmarks of the demographics of the 2016 voting electorate remain unavailable. However, some early signs suggest that the unique turnout patterns of 2016 may have contributed to polling error.

The composition of the 2016 electorate was different from four years earlier in ways that advantaged Trump and disadvantaged Clinton. That much is now clear from both county vote totals and the analysis emerging from official records of individual voter behavior. As of this writing, only a subset of states have released updates to their official voter registration lists to include information about 2016 turnout. Also, the biennial update to the U.S. Census Current Population Survey Voting and Registration Supplement, the U.S. Census survey designed to measure voter turnout and voter characteristics, will not be released until May or June 2017.

However, in those states that have updated their registered voter files (the term commonly used to describe official registration lists) the emerging pattern is consistent on several key points. The most critical is that, in those states examined so far, turnout appears to have been higher among Republicans than among Democrats compared to 2012.

- In Pennsylvania, for example, Catalist, a data firm aligned with the Democratic Party, estimates that those who voted in 2012 but not 2016 were more likely to be registered Democrats than those who voted in 2016. They identify 1.1 million registered voters in Pennsylvania who voted in 2012 but not in 2016; 52 percent were registered Democrats, 35 percent registered Republicans. Those who turned out in 2016 were 47 percent Democratic, 41 percent Republican.

- In both Pennsylvania and North Carolina, Nate Cohn (2017) examined official voter files and found the turnout of African Americans in 2016 to be significantly lower than in 2008 or 2012. Cohn also found the composition of the electorates in Florida, North Carolina and Pennsylvania to be slightly less Democratic than had been predicted by his statistical models.

- In Colorado, TargetSmart, another Democratic-aligned data firm, found those who voted in 2012 but not 2016 tended to be younger and were very slightly more Republican within each age group.

- In Iowa, a state with one of the largest vote swings from Obama to Trump, Republican aligned data analyst Patrick Ruffini examined voter files and found that turnout among registered Democrats was down, while turnout among Republicans roughly matched levels forecast by turnout models based mostly on whether voters had cast ballots in 2012.

- In North Carolina, where voters are classified by race in accordance with the Voting Rights Act, Ruffini found a “massive turnout gap on racial lines.” African Americans voted at lower rates than predicted by their 2012 performance, while white turnout

exceeded expectations based on the 2012 benchmark. The result was a "significant gap" in partisan turnout, with registered Republicans exceeding projections based on 2012 and Democrats falling short.

- In Ohio, Ruffini found similar differences on racial lines (although in Ohio the racial classifications were based on statistical modeling, not official records). Here too, a gap emerged by partisanship, with voters with a history of voting in Democratic primaries falling slightly farther below their predicted behavior, based on 2012 benchmarks, than Republicans.

Patterns consistent with these results are also apparent in the certified county-level vote counts across the country. Nationwide, the total vote cast for President in 2016 grew by just under 6 percentage points as compared to 2012. As shown in Figure 12, that growth was slightly larger (7 percent) in counties where Barack Obama received 45 percent of the vote or less in 2012 and slightly lower (5 percent) in counties where Obama had received 55 percent or higher.

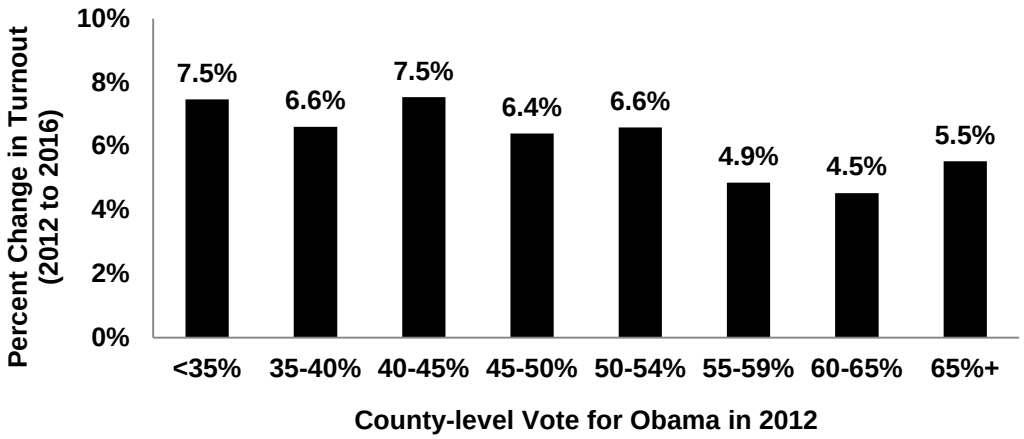


Figure 12. Change in Turnout (2012 to 2016) by County Nationwide by Vote Share for Barack Obama in 2012. Source – County-level vote data come from uselectionatlas.org.

As report co-author Douglas Rivers found, the contrast between the growth of the vote in Republican counties and Democratic counties was especially pronounced in several critical Midwestern battlegrounds, including Michigan, Wisconsin and Ohio (Leonhardt 2016). Further analysis shows similar patterns in many Southern and border states, such as Alabama, Arkansas, Mississippi, Missouri, South Carolina and Tennessee and Plains States such as Nebraska and North and South Dakota. Of course, these comparisons may misstate underlying turnout differences due to population growth that is typically greatest in suburban counties. So again, while a complete accounting of the turnout of 2016 will benefit from a fuller analysis of voting records and the Bureau of Labor Statistic's Current Population Survey (CPS) Voting and Registration Supplement, the early evidence points to a shift in turnout that helped Trump and hurt Clinton.

For the purposes of this report, the question is whether pollster assumptions about turnout contributed to polling errors in 2016, particularly at the state level. Some pollsters lean heavily on the assumption that the past election is the best possible model of the coming election, and

some do not (Appendix A.J discusses various approaches to likely voter modeling). So knowing that the demographics of the 2016 turnout differed from 2012 in some respect does not automatically translate into a polling surprise. Thus, assumptions aside, some polls may have captured the dynamics of turnout in 2016 better than others.

3.5.2 Measuring How Accurately Polls Predicted Turnout

Aside from the accuracy of the horserace forecast, we can measure how accurately polls captured the actual turnout in three ways, at least in theory. The first and most direct involves validating vote history for each respondent. Two other methods are more indirect, comparing the geographic and demographic compositions of the poll samples to what we know about the actual turnout. Each of these has strengths and weaknesses. Researchers continuing this line of research will need more time and data to better assess and quantify the degree to which the 2016 polls accurately captured turnout – given the level of disclosure and available raw data – much may remain unknown. But early measurements suggest that the changes in turnout that favored Trump may have contributed to polling error, perhaps more in some states than others.

Validation

The most direct and straightforward way to evaluate how well a poll predicted turnout is to validate which respondents voted and which did not. Such an exercise can shed light on how accurately the pollster's methods identified likely voters, and on whether either non-voters included in the sample, or actual voters left out, contributed to any error in estimating the ultimate result. Unfortunately, a full validation is neither easy nor feasible for the vast majority of public polls. Polls conducted by telephone rarely attempt to ask and record the full name and street address of every respondent – the information necessary to attempt anything approaching a complete match to official records. Phone numbers are sometimes used for validation matching purposes, but at best, these allow for a match at the household level and only when phone numbers have been matched to voting records.

Practically speaking, the surveys most able to validate turnout are those that sampled directly from voter lists and interviewed specific voters, by name, allowing for a full match to voter file data. In such instances, the match back to vote history records is relatively straightforward, once the voter files have been updated to include 2016 turnout data. Unfortunately, very few surveys whose results were made public in 2016 sampled from voter lists in a way that readily facilitates validation, though a few were. One example involves a series of polls conducted for the College of William & Mary by Target Smart, a Democratic-aligned data firm. A post-election analysis found that their respondents who actually voted were more likely to support Trump than respondents who did not vote. Excluding the validated non-voters moved their estimate of Trump's percentage of the two-party (Trump and Clinton) vote from 51 to 53 percent. Trump received 54 percent of the two-party vote in Ohio.

Geographic Composition

A second method of evaluating how well polls matched turnout is through their demographic composition. This approach has some important advantages. We now have complete, certified vote counts for every county in the United States that can tell us how the total vote in 2016 compared to 2012. For polls that ask respondents to identify their county of residence, we can array the counties in each state by their past voting and assess whether poll samples over-stated

the contribution of heavily Democratic or heavily Republican counties and whether correcting any such discrepancy would reduce polling error.

The main disadvantage is that there is far more to the makeup of the electorate than county-level composition. A lot can remain wrong about the selection of likely voters even if the county-level geographic composition is spot on. As noted above, turnout nationwide typically grew more in heavily Republican counties than in heavily Democratic counties. While these turnout patterns were likely a factor in Trump's very narrow margins in states like Michigan and Wisconsin, the *geographic* shifts *alone* would rarely have been enough to introduce significant *survey error* for those pollsters who weighted by geography to match the distribution of the vote in 2012.

Consider Michigan, a state where turnout clearly increased most in the most heavily Republican counties. The total votes cast for president fell by 2 percent in the two Democratic counties, Wayne and Washtenaw, where Barack Obama received over 65 percent of the vote in 2012, but tended to increase elsewhere, with the biggest gains (over 8 percent) coming in the heavily Republican counties where Obama's vote was 35 percent or lower.

Table 12. County-level Change in Turnout in Michigan

Counties Where Obama's 2012 Vote Percentage Was….	Percent Change	2012 Composition	2016 Composition
<35%	8.5%	2.9%	3.1%
35-40%	6.4%	5.2%	5.5%
40-45%	3.9%	10.2%	10.4%
45-50%	3.4%	19.1%	19.3%
50-54%	2.8%	27.6%	27.8%
55-59%	1.0%	7.0%	6.9%
60-65%	-0.3%	7.0%	6.8%
65%+	-2.3%	21.0%	20.2%
		100.0%	100.0%

Source – County-level vote data come from uselectionatlas.org.

Nevertheless, these changes altered the geographic *composition* of Michigan's electorate only slightly. The percentage of the state's total vote coming from those two most heavily Democratic counties fell, but only slightly, from 21.0 to 20.2 percent. Meanwhile, the share coming from the more Republican counties where Obama's vote total had been under 45 percent grew, but again by less than a single percentage point (from 18.3 to 18.9 percent). Had a pollster weighted by these strata using the 2012 composition, the resulting error in vote share estimates would be minimal – often not changing a candidate's share of the vote by enough to round up or down by a single digit.

Demographic Composition

An unexpected change in the demographics of the voting electorate is more likely to lead to survey error, particularly if the change led to a significant miss in terms of race, ethnicity, age or

gender. Race and ethnicity would have an especially powerful impact, given that the NEP exit poll and most pre-election surveys showed Clinton winning nearly 90 percent of African Americans, better than 60 percent of Latinos and Asians, but under 40 percent of white voters. Thus, an over-statement of the African-American share of the vote would increase Clinton's margin over Trump on a nearly one-to-one basis – a 1 percentage over-statement of African-Americans would increase her margin by nearly a full point.

However, many obstacles prevent a full evaluation of how well polls matched the demographics of 2016 turnout. First, many of the public polls provide little or no disclosure of their weighted composition in terms of gender, age, race and ethnicity. Second, the most important benchmarks of voter demographics often conflict with respect to age and race. Michael McDonald (2007) found that official voter files and the CPS were largely in agreement with respect to the demographic profile of selected state electorates, but that NEP exit poll estimates tended to be "younger and composed of more minorities." The consistency of voter files with CPS, McDonald concluded, "lends validity to their information and should comfort scholars as they investigate these data." More recently, Nate Cohn (2016b) also compared the three sources and reached a similar conclusion.

As of this writing, the most credible sources of information about the demographic composition of the 2016 voting electorate are mostly unavailable. The CPS Voting Supplement results will not be published until mid-2017, and only a handful of states so far have updated their voter files with 2016 vote history data. Even then, the voter files include the voter's race only in a small handful of southern states covered by the Voting Rights Act.

3.5.3 Assessing the Effect of Weighting and Likely Voter Modeling Using Microdata

One possible scenario is that the raw data collected by pollsters in key battleground states was relatively accurate but well-intentioned demographic adjustments or likely voter modeling led the polls astray. Figure 13 shows the signed error on the presidential vote margin for polls in four key battleground states. For each poll, the weighted likely voter (LV) estimate is shown in black, the weighted registered voter (RV) estimate is shown in gray, and the unweighted RV estimate is shown in white.[18] The higher the value, the more the estimate over-stated support for Clinton, relative to the election outcome. This analysis was only possible with polls for which the committee had microdata. It is also important to note that several of the polls included in the analysis were fielded more than two weeks out from Election Day and not intended to be a final projection of the contest.

[18] The Monmouth microdatasets did not have a variable to distinguish LVs from all RVs, so no weighted RV estimates are presented for those polls.

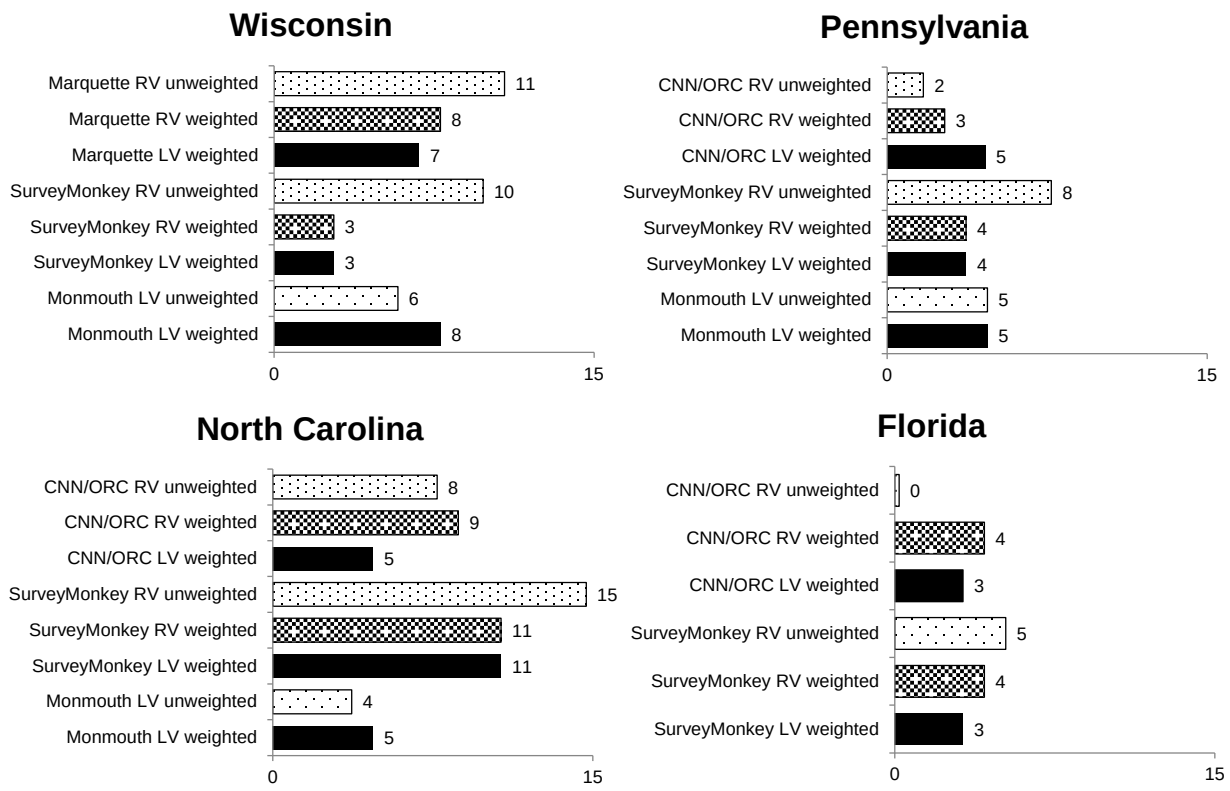


Figure 13. Signed Error on 2016 Presidential Vote Margin by Poll, Level of Modeling and State. Notes – The Marquette University poll was fielded October 27-30, 2016. The SurveyMonkey polls were all fielded November 1-7, 2016. The Monmouth University polls were fielded October 15-18, 2016 (Wisconsin), October 20-23, 2016 (North Carolina) and October 29-November 1, 2016 (Pennsylvania). The CNN/ORC polls were fielded October 10-15, 2016 (North Carolina) and October 27-November 1, 2016 (Pennsylvania and Florida).

The results point to inconsistent effects across polls from weighting and likely voter modeling. In the SurveyMonkey polls, conducted online with opt-in sample, the weighting clearly helped to improve accuracy. For example, in Wisconsin, the error on the margin was 10 points in the unweighted data but just 3 points in the weighted data. Analysis of the poll's demographic profile shows the weighting was primarily correcting an imbalance on education not race/ethnicity or geography. As discussed in Section 3.4, education was strongly, linearly associated with presidential vote this year, so the weighting was critical for reducing error in the SurveyMonkey data.

The pattern for CNN/ORC polls, conducted by live telephone with RDD sample, was quite different. CNN/ORC's unweighted data was basically spot on the margin in Florida and quite close in Pennsylvania. In those states, weighting and likely voter modeling increased the signed error several percentage points, making the final figures too Democratic. Non-Hispanic blacks comprised 10 percent of CNN/ORC's unweighted RV sample in Florida but 14 percent of the weighted LV sample. Since the poll had blacks favoring Clinton by 92 points, that adjustment (which probably would have improved accuracy in an election with higher Democratic turnout) had the net effect of pushing the published margin farther from the vote outcome. The CNN/ORC data in Pennsylvania tell the same story. In North Carolina, by contrast, likely voter modeling improved the CNN/ORC poll.

Figure 13 shows that in Wisconsin, statistically adjusting the data slightly helped the Marquette University poll and slightly hurt the Monmouth University poll. In both cases, weighting and/or likely voter modeling had virtually no effect on the race distribution. The Marquette poll weighting, however, noticeably reduced the influence of college graduates (by 12 percentage points), while the Monmouth weighting did not.

This analysis demonstrates that different pollsters made different assumptions about the education and race/ethnicity profile of the voting electorate in 2016. In these battleground states, weighting down college graduates helped to improve accuracy while weighting up non-Hispanic blacks appears to have reduced accuracy. This result comports with reporting of turnout patterns in the election.

Overall, the analysis shows that post-survey statistical adjustment reduced error in these polls, specifically reduced over-estimation of Clinton support – by about 2 percentage points. Unfortunately, the adjustment did not succeed in reducing the error to zero, but the direction of the effect indicates that pollsters were conscious of the fact that their data needed adjustment and they were, in most cases, making the adjustments in the proper direction.

Statistical adjustment, in itself, does not appear to be an important cause of error in the polls broadly speaking. One could argue that *more* adjustment was needed in some polls, but that means that the root problem was something else, not the adjustment. Of the 11 polls in this analysis, there is only one poll, CNN/ORC in Florida, where statistical adjustment appears to be the reason for the projection miss, but that would completely ignore the evidence for late movement in the Florida contest during the final week (interviewing finished November 1st). In most other polls examined, statistical adjustment is a reason why the projection miss was not worse.

This analysis is based on a limited number of pollsters in a limited number of states. No IVR pollsters agreed to provide microdata to the committee, so no IVR polling is reflected here. Among RDD and online opt-in polls, CNN/ORC and SurveyMonkey were relatively prolific in 2016, and their weighting protocols look to be reasonably representative (if on the more rigorous side) of most live telephone RDD and opt-in online polls in the 2016 general election.

3.6 Ballot Order Effects

Another possible explanation for polling error in 2016 is what is known as a ballot order effect (BBC News 2017; Pasek 2016; Gelman 2017). Political methodologists have documented a small but non-trivial bias in favor of candidates listed first on election ballots (e.g., Ho and Imai 2008; Miller and Krosnick 1998; Pasek et al. 2014). While pollsters typically eliminate this effect in the polls by randomizing the order of the candidate names presented, most states do not. Instead, they list presidential candidates in the same order in every county and every precinct. Only a minority use a rotation or randomization process to avoid the primacy effect of voters who are only weakly attached to their choice voting for the first candidate on the ballot.

In assessing the potential effect of ballot order on the performance of the polls, the first thing to note is the size of the effect. One study found no detectable effect from ballot order on the vote

for major party presidential candidates (Ho and Imai 2008), but a more recent study reported an improvement of 0.3 percentage points (Pasek et al. 2014) for the candidate listed first. An effect of this size is large enough to have influenced the outcome in a state like Michigan, where Trump won by 0.2 percentage points and was listed first on the ballot. Ballot order is also likely to have helped Trump in Wisconsin and Florida, where he was also listed first statewide, but it is not at all clear that ballot order tipped the outcome in those contests given that Trump won them by larger margins (0.8 and 1.2 percentage points, respectively).

While an effect of about 0.3 percentage points is a large enough to potentially explain the outcome in at least one state, it is not large enough to explain a meaningful amount of the error in 2016 pre-election polls. As noted in section 2.2, the average absolute error was 3.6 points for battleground state polls and 6.4 points for non-battleground state polls. Put differently, the average error in the state polls was roughly 10 to 20 times larger than the estimated size of the ballot order effect.

Even though this effect was clearly a minor one for pollsters, the committee conducted analysis with the available microdata. Neither the October Pew Research Center poll (conduct via live phone with RDD sample) nor YouGov polling (conducted via an online opt-in panel) showed a perceptible effect from candidate name order. Both surveys randomized candidate name order, and the estimated share endorsing Trump was virtually the same when his name was presented first versus second. This analysis does not rule out the possibility that ballot order influenced the outcome in states like Michigan, but it does bolster the conclusion that ballot order was not an important explanation as to why polls under-estimated support for Trump in the general election.

4. POLLING AND PROBABILISTIC FORECASTING

There is nothing new about the public fascination with either polls or predictions. They have been part of both the cultural and political chatter for decades. At times, polling has seemed incomparably clairvoyant and, at other times, miserably misdirected. It is neither. Public opinion polling is a method of measurement. When scientifically based, it has the capability to provide insight into the opinions, attitudes and behaviors of the many by connecting with the few. Its results are estimates, within a range, at a particular point in time. Any one poll is a static measure. The benefit of public polls is to provide a narrative about the status of issues and elections for the public, the press, and the polity.

In contrast, many polling aggregators are prognosticators. Among these forecasters are a few scholars of polling, but many are masters of other fields. They come together as data scientists or data journalists charged with accumulating numbers from a multitude of sources to identify patterns and predict future events. In politics, that means innovating experimental models to predict elections.

Although one source of data for most of the polling aggregators is public polling, each one has its own recipe for deconstructing the political stew. These recipes vary considerably. At one extreme, *RealClearPolitics* takes an average of public polls nationally to gauge the popular vote, and as available, within each state to capture candidate strength. *RealClearPolitics* then characterizes a state on a continuum from strong Democrat to strong Republican. On the other

extreme is *FiveThirtyEight* which offers several models including a forecast of each candidate's chance of winning. These are based on poll averages, adjustments to polls based on several factors, other demographic and/or economic data, and election simulations to account for statistical error and uncertainty.

What these and other sources have in common is their enormous reach. Consider these two aggregators, *RealClearPolitics* and *FiveThirtyEight*. The total number of visits to these sites in the month of October was nearly 200 million. This does not include the avalanche of comments and shares on social media, or the voluminous mentions in news media and conversation. According to the media monitoring service *Meltwater*, an estimate of the advertising value equivalency of their combined media mentions was $1.025 billion in *just the final week* of the campaign. Not included in these calculations are additional election prognosticators such as *HuffPost Pollster, The New York Times' Upshot, PredictWise, Princeton Election Consortium, Daily Kos, The Cook Political Report, Roll Call's Inside Elections,* and *Sabato's Crystal Ball*.

Despite the varied assumptions in constructing their estimates (polls, factors for adjusting poll results, the independence of errors across measures or geography, previous election results, demographic and/or economic data), they predicted a strong chance of a Clinton win often down to the decimal point (suggesting to the audience far more precision than warranted). Collectively, the probabilistic modelers declared that Clinton's likelihood of winning was about 90 percent, with estimates ranging from 71 to over 99 percent (Katz 2016). Now that the dust has settled, it is clear that most of those models under-stated how competitive the race actually was. Summarizing the state-based polling, RealClearPolitics' Sean Trende noted (2016), "Trump stood just one state short of winning the presidency." No runaway for Clinton based only on public polls should have been expected. Nate Silver of FiveThirtyEight, which had the lowest probability of a Clinton win, concurred. He tweeted, "… it's irresponsible to blame the polls for the over-confidence in Clinton's chances. They showed a competitive, uncertain race."

Another reason greater caution is needed in predicting the Electoral College is the uneven quality of state-level polls, including the haphazard nature of who conducts them and when. As discussed above, the universe of state-level polls is qualitatively different and generally speaking less rigorous than national polls. Sample sizes tend to be smaller and weighting adjustments tend to be less comprehensive, among other factors. As Frank Newport remarked (2016), "To the degree that organizations want to predict the Electoral College, they are going to have to find ways to finance or encourage larger-sample, higher-quality state polls, rather than relying on the haphazard polls that happen to be conducted in the various states."

One final concern with both polling and probability forecasts is their potential effect on the election outcome itself. Does a poll showing a large lead in a state make that state's citizens less likely to bother to vote? Does a probabilistic forecast that a candidate has more than a 99 percent chance of winning have a similar effect? Does the latter have a stronger effect? However well-intentioned and scientifically motivated these estimates may be, their effects on human behavior are not well understood. While answering these questions is beyond the scope of this report, the events of 2016 suggest they deserve serious attention.

To be clear, we believe that polling serves an important function in a democracy and should not be curtailed. At the same time, the massive backlash and distrust of polls in the wake of the election makes clear that how polls, poll-based forecasts and the limitations of both are communicated to the public is sorely in need of improvement. Low-hanging fruit here include steps like reporting more accurate margins of error (e.g., Rothschild and Goel 2016); reporting margins of error for the margin between the top two candidates (which is often what gets attention) not just a hypothetical estimate of a candidate with 50% support; better explaining the implications of the margin of error; avoiding reporting decimal points on estimates that, if lucky, come within several points of reality; and better accounting for the possibility of correlated errors (Silver 2016a).

5. CONCLUSIONS

The committee, commissioned by AAPOR, conducted an extensive investigation of the performance of pre-election polls in 2016. While the general public reaction was that "the polls failed," we found the reality to be more complex – a position held by a number of industry experts (Newport 2016; Silver 2016a; Trende 2016). Some polls, indeed, had large, problematic errors, but many polls did not. Critically, the reasons for the polling errors are no longer a mystery. We found evidence for several factors that led polls to under-state support for Trump relative to the election outcomes in battleground states. Here are the conclusions, in summary:

The national polls were generally correct and accurate by historical standards.
The national polls indicated, on average, that Clinton would win the national popular vote by about 3 percentage points. They were reasonably accurate; she ultimately won the national popular vote by 2.1 percentage points. Putting it another away, using the two metrics comparing poll performance across time, the average absolute error in the polls was also 2.2 percentage points. The average signed error was 1.3 percentage points. Both are good performances by historic standards. This performance was not the result of two large errors canceling (under-estimation of Trump support in heavily working class white states and over-estimation of his support his liberal-leaning states with sizable Hispanic populations).

The state polls had a historically bad year in terms of forecasting the state outcomes.
In contrast, the state-level polls, had an average absolute error of 5.1 percentage points and a signed error of 3.0 points. This is the largest average absolute error in the elections starting in 2000. And the overall signed error was in the direction of over-estimating support for Clinton and under-estimating support for Trump.

In key states, the polls' under-estimation of Trump's support was pivotal.
In more competitive states, the errors by the state polls were lower than those for polls in the non-competitive states. But in the competitive states the under-estimation of Trump's support led to incorrect conclusions, especially in two states in the Upper Midwest. Polls showed Clinton leading, if narrowly, in Pennsylvania, Michigan and Wisconsin. Those leads fed predictions that the Democratic *Blue Wall* would hold. Come Election Day, however, Trump eked out victories in all three.

Reasons for under-estimating Trump's support: Late changes in voter choices.

There are a number of reasons as to why polls under-estimated support for Trump in various polls. There is evidence of real late change in voter preferences in Trump's favor in the last week or so of the campaign, especially in the states where Trump won narrowly.

Reasons for under-estimating Trump's support: Pollsters' failure to weight by education. Education was strongly correlated with presidential vote in key states: that is, voters with higher education levels were more likely to vote for Clinton. Yet some pollsters – especially state-level pollsters – did not adjust for education in their weighting even though college graduates were over-represented in their surveys. This led to an under-estimation of support for Trump.

Reasons for under-estimating Trump's support: Little backing for *Shy Trump* hypothesis. During both the primaries and the general election, some postulated that Trump supporters might be unwilling to tell a live interviewer their intentions, on the theory that backing Trump was not a socially acceptable view. A number of Trump voters who participated in pre-election polls did not reveal themselves as Trump voters until callback studies conducted after the election (and they outnumbered late-revealing Clinton voters), which could be attributable to either late deciding or misreporting (*Shy Trump*) in the pre-election poll. A number of other tests for the *Shy Trump* hypothesis yielded no evidence to support it, including differences between polls with live interviewers and those with no live interviewers.

Reasons for under-estimating Trump's support: Nonresponse bias.
Response rates in telephone polls with live interviewers continue to decline, with response rates even lower for other methodologies. Thus, some have argued that the national polls must have benefited from offsetting errors in areas with high Trump support and those with high Clinton support. The evidence does not support that nor does it support that pro-Trump areas were systematically under-represented in national or state polls.

Reasons for under-estimating Trump's support: Turnout changes and flaws in models.
Turnout patterns changed between 2012 and 2016 that could have led to mistakes in likely voter models used by pollsters. The best data sources for examining this had not been released in time for this analysis: the Current Population Survey Voting and Registration Supplement or updated registered voter files from across the country. These changes in turnout and mistakes in likely voter models could have contributed to some polling errors, though the analysis that we were able to conduct examining the impact of likely voter modeling showed generally small and inconsistent effects.

Reasons for under-estimating Trump's support: Ballot order in some states.
Ballot order effects may have played a role in some state contests, but they don't go far in explaining the polling errors. Scholars have noted that state election rules led to Trump's name appearing above Clinton's on all ballots in several key states that Trump won narrowly (Michigan, Wisconsin and Florida). In the context of polling errors that were substantially larger that estimated ballot order effects, ballot order represents at best only a minor reason for polling problems.

Presidential primary polls generally performed on par relative to past elections.

The 2016 pre-election estimates in the Republican and Democratic primaries were not perfect, but the misses were normal in scope and magnitude. The vast majority of primary polls predicted the right winner. When polls did badly miss the mark, it tended to be in contests where Clinton or Trump was the runner-up.

The poll aggregators and estimators.
Aggregations of poll results and projections of election results had a difficult year in 2016. They helped crystalized the erroneous belief that Clinton was a shoo-in for president. While a similar criticism can be leveled against polls – i.e., they can indicate an election is uncompetitive, perhaps reducing some people's motivation to vote – polls and forecasting models are not the one and the same. Pollsters and astute reporters are often careful to describe their findings as a snapshot in time, measuring public opinion at the time they are fielded. Forecasting models do something different – attempt to predict a future event. As the 2016 election proved, that can be a fraught exercise. Caution and humility seem to be in order for pollsters and those who use polls.

There is no partisan favoritism in U.S. polling.
In 2016 national and state-level polls tended to under-estimate support for Trump, the Republican nominee. In 2012 and 2000, however, general election polls clearly tended to under-estimate support for the Democratic presidential candidates. The trend lines for both national polls and state-level polls show that – for any given election – whether the polls tend to miss in the Republican direction or the Democratic direction is essentially random.

Election polls are not all polls.
The difficulties for election polls in 2016 are not an indictment on all of survey research or even all of polling. The performance of election polls is not a good indicator of the quality of surveys in general. Election polls are unique among surveys in that they not only have to field a representative sample of the public but they also have to correctly model who among that sample will actually vote in the election. The second task presents substantial challenges that a non-election poll simply does not have. It is, therefore, a mistake to observe errors in an unusual election such as 2016, and conclude that all polls are broken. That is simply not accurate and it is refuted by a large body of research that shows many surveys are still able to produce valuable, reasonably accurate information about the attitudes and experiences of the U.S. public (Ansolabehere and Schaffner 2014; Bloom and Pearson 2008; Dutwin 2016; Keeter et al. 2006; Kennedy, Wojcik and Lazer 2017; Pew Research Center 2012).

The analysis and insights in this report were made possible by members of the polling community committed to advancing the science of public opinion measurement. We are grateful to the numerous professionals who answered our calls or emails and shared their knowledge. We hope that in future efforts more organizations, especially those representing the IVR sector of the field, will demonstrate this same commitment to scientific advancement generally and the AAPOR Transparency Initiative specifically. We have little doubt that broader cooperation would result in even more robust and important findings.

Correction: May 11, 2017
An earlier version of this report misstated the average signed error for national polls in the 2016 general election as 0.7 percentage points. This version reports the corrected value of 1.3 percentage points.

References

Agiesta, J. (2016), "The Polling Shapshot: Heading to New Hampshire, Poll Overload," CNN.com, February 8, 2016. Retrieved from http://www.cnn.com/2016/02/08/politics/new-hampshire-polling-snapshot/.

American Association for Public Opinion Research (AAPOR) (2009), "An Evaluation of the Methodology of the 2008 Pre-Election Primary Polls," Retrieved from http://aapor.org/uploads/AAPOR_Rept_FINAL-Rev-4-13-09.pdf.

Ansolabehere, S. and B.F. Schaffner (2014), "Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison," *Political Analysis* 22(3): 285-303.

Battaglia, M.P., M.R. Frankel and M.W. Link (2008), "Improving Standard Poststratification Techniques For Random-Digit-Dialing Telephone Surveys," *Survey Research Methods*, 2(1): 11-10.

BBC News (2017), "Did Trump win because his name came first in key states?" BBC News, February 25, 2017. Retrieved from http://www.bbc.com/news/magazine-39082465.

Blake, A. (2016), "How America decided, at the last moment, to elect Donald Trump" Washington Post, November 17, 2016.

Bloom, J. and J. Pearson (2008), "Reliable Compared to What? A Probability-Theory Based Test of the Reliability of Election Polls," *In Elections and Exit Polling*, Scheuren, F. and W. Alvey, Eds. Wiley and Sons: New York.

Blumberg, S. and J.V. Luke (2016), "Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, January–June 2016." National Center for Health Statistics. Available from: http://www.cdc.gov/nchs/nhis.htm.

Blumenthal, M. (2016), "The Latest Data and Methodological Information on How SurveyMonkey Measures Shifts in Voter Sentiment" Retrieved from https://blog.electiontracking.surveymonkey.com/2016/12/22/looking-back-at-2016-what-weve-learned-so-far/.

Bodor, T. (2012), "The Issue of Timing and Opinion Congruity in Spiral of Silence Research: Why Does Research Suggest Limited Empirical Support for the Theory? International Journal of Public Opinion Research Vol. 24 No. 3, 269-286.

Bradner, E. (2016), "Clinton Leads trump, Two New Polls Show." Retrieved from http://www.cnn.com/2016/10/16/politics/hillary-clinton-donald-trump-presidential-polls/

Byers, D. (2016), "How Politicians, Pollsters and Media Missed Trump's Groundswell." Money.CNN.com. Retrieved from http://money.cnn.com/2016/11/09/media/polling-media-missed-trump/.

Cassino, D. (2016), "How Today's Political Polling Works," Harvard Business Review, August 1, 2016. Retrieved from https://hbr.org/2016/08/how-todays-political-polling-works.

Chang, L and J.A. Krosnick (2009), "National Surveys Via Rdd Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality" *Public Opinion Quarterly* 73(4): 641-678.

Cho, S., J. Cohn, J. Chen, M. Blumenthal, and L. Wronski (2016), "SurveyMonkey Election Tracking: Learnings from 1 Million Interviews" Presented at the PAPOR Annual Conference.
Chozick, A. (2016), "Hillary Clinton Blames F.B.I. Director for Eelction Loss." New York Times, November 12, 2016. Retrieved from https://www.nytimes.com/2016/11/13/us/politics/hillary-clinton-james-comey.html.
Cillizza, C. (2016), "Winners and Losers from the 2016 Election" Washignton Post, November 9, 2016. Retrieved from https://www.washingtonpost.com/news/the-fix/wp/2016/11/09/winners-and-losers-from-the-2016-election/?tid=a_inl&utm_term=.cbb00e865a74.
Clement, S. (2016), "The 2016 National Polls Are Looking Less Wrong after Final Election Tallies," Washington Post, February 6, 2017. Retrieved from https://www.washingtonpost.com/news/the-fix/wp/2016/11/10/how-much-did-polls-miss-the-mark-on-trump-and-why/?utm_term=.4bb2e462d651.
Clinton, J.D. and S. Rogers (2012), " Robo-polls: Taking Cues from Traditional Sources?" *PS: Political Science & Politics* 46(2): 333-337.
Cohn, N. (2014), "Two Polls That Highlight the Challenges of Polling" New York Times, October 14, 2014, Retrieved from https://www.nytimes.com/2014/10/15/upshot/two-polls-that-highlight-the-challenges-of-polling.html.
Cohn, N. (2016a), "Is Traditional Polling Underselling Donald Trump's True Strength?" New York Times, May 17, 2016. Retrieved from https://www.nytimes.com/2016/05/18/upshot/is-traditional-polling-underselling-donald-trumps-true-strength.html?_r=1.
Cohn, N. (2016b), "There Are More White Voters Than People Think. That's Good News for Trump." New York Times, June 9, 2016. Retrieved from https://www.nytimes.com/2016/06/10/upshot/there-are-more-white-voters-than-people-think-thats-good-news-for-trump.html?_r=0.
Cohn, N, J. Katz and K. Quealy (2016), "Putting the Polling Miss of the 2016 Election in Perspective," New York Times, November 12, 2016, Retrieved from https://www.nytimes.com/interactive/2016/11/13/upshot/putting-the-polling-miss-of-2016-in-perspective.html.
Cohn, N. (2017), "A 2016 Review: Turnout Wasn't the Driver of Clinton's Defeat," New York Times, March 28, 2017, Retrieved from https://www.nytimes.com/2017/03/28/upshot/a-2016-review-turnout-wasnt-the-driver-of-clintons-defeat.html?_r=0.
Collins, E. (2016), "Poll: Clinto, Trump Most Unfavorable Canddiates Ever." USA Today August 31, 2016. Retrieved from http://www.usatoday.com/story/news/politics/onpolitics/2016/08/31/poll-clinton-trump-most-unfavorable-candidates-ever/89644296/.
Crespi, I. (1988), *Sources of Accuracy and Error in Pre-Election Polling*. New York: Sage.
Director of National Intelligence (2017), "Background to "Assessing Russian Activities and Intentions in Recent US Elections: The Analytic Process and Cyber Incident Attribution," Retrieved from https://www.dni.gov/files/documents/ICA_2017_01.pdf.
Donovan, L. (2016), "How Democrats Misread Their Baseline – and their Base," National Review, November 16, 2016, Retrieved from http://www.nationalreview.com/article/442224/donald-trumps-victory-democrats-blue-wall-crumbles.

Dropp, K. (2016), "How We Conducted Our 'Shy Trumper' Study," Retrieved from https://morningconsult.com/2016/11/03/shy-trump-social-desirability-undercover-voter-study/.
Dutwin, D. (2016), "Political Polling Isn't Dead Just Yet," Washington Post, January 28, 2016. Retrieved from https://www.washingtonpost.com/news/the-fix/wp/2016/01/28/reports-of-the-demise-of-polls-have-been-greatly-exaggerated/?utm_term=.26df9287ea92.
Easley C. (2016a), "50-State Snapshot: Clinton Beats Trump, but It's Close," MorningConsult, July 14, 2016. Retrieved from https://morningconsult.com/trump-vs-clinton-conventions-electoral-college-map/.
Easley, J. (2016b), "Pollsters Suffer Huge Embarrassment." The Hill, November 9, 2016. Retrieved from http://thehill.com/blogs/ballot-box/presidential-races/305133-pollsters-suffer-huge-embarrassment.
Enten, H. (2012), "The Other 2012 Election Contest: Which Pollster and Polling Method Will Win?" The Guardian, October 31, 2012. Retrieved from https://www.theguardian.com/commentisfree/2012/oct/31/other-2012-election-contest-pollster-polling.
Enten, H. (2016), "'Shy' Voters Probably Aren't Why The Polls Missed Trump," FiveThiryEight.com, November 16, 2016. Retrieved from https://fivethirtyeight.com/features/shy-voters-probably-arent-why-the-polls-missed-trump/.
Forsberg, O. (2016), "Knox Professo Sheds Light on Polling and 2016 Election," November 11, 2016, Retrieved from https://www.knox.edu/news/knox-professor-sheds-light-on-polling-and-election-2016.
Gelman, A. (2017), "Did Trump Win Because His Name Was First on the Ballot?" Slate, February 28, 2017. Retrieved from http://www.slate.com/articles/health_and_science/science/2017/02/trump_s_win_probably_wasn_t_due_to_where_his_name_was_on_the_ballot.html.
Gelman, A., S. Goel, D. Rivers and D. Rothschild (2016), "The Mythical Swing Voter," *Quarterly Journal of Political Science*, 11(1) 103-130.
Goldmacher, S. (2016), "Clinton Looks Poised to Lock it Up," Politico, November 11, 2016, Retrieved from http://www.politico.com/story/2016/11/hillary-clinton-donald-trump-countdown-230856.
Gore, D. (2016), "Presidents Winning Without Popular Vote," FactCheck.org, December 23, 2016, Retrieved from http://www.factcheck.org/2008/03/presidents-winning-without-popular-vote/.
Ho, D.E. and K. Imai (2008), "Estimating Causal Effects of Ballot Order from a Randomized Natural Experiment: The California Alphabet Lottery, 1978-2002," *Public Opinion Quarterly* 72(2): 216-240.
Jackson, N. (2016), "Here's Why HuffPost Is Dropping Polls That Rely Only On Landlines" Huffington Post, August 1, 2016. Retrieved from http://www.huffingtonpost.com/entry/landline-only-polls-huffpost-pollster_us_579f9b2ae4b08a8e8b5ee65e.
Jacobs, J., and B. House (2016), "Trump Says He Expected to Lose Election Because of Poll Results," Bloomberg.com, December 13, 2016. Retrieved from https://www.bloomberg.com/politics/articles/2016-12-14/trump-says-he-expected-to-lose-election-because-of-poll-results.

Jacobson, L. (2016), "Will Online Polls Revolutionize Elections?" Governing.com, February, 1, 2016, Retrieved from http://www.governing.com/topics/elections/gov-polling-american-panel-survey.html.
Jennings, W. and C. Wlezien (2016), "The Timeline of Elections: A Comparative Perspective," *American Journal of Political Science* 60(1): 219-233.
Kang, C. (2016), "Fake News Onslaught Targets Pizzeria as Nest of Child-Trafficking," *New York Times*, November 21, 2016.
Katz, J. (2016), "Who Will Be President?" New York Times. November 8, 2016. Retrieved from https://www.nytimes.com/interactive/2016/upshot/presidential-polls-forecast.html.
Keeter, S., C. Kennedy, M. Dimock, J. Best, and P. Craighill (2006), "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey." *Public Opinion Quarterly*, Vol 70(5): 759-779.
Kennedy, B. S. Wojcik, and D. Lazer (2017); "Improving Election Prediction Internationally," *Science* 355(6324): 515-520.
Kohut, A. (2006), "Andrew Kohut Interview." Pollster.com, October 23, 2006. Retrieved from http://www.pollster.com/blogs/andrew_kohut_interview.html.
Leonhardt, D. (2016), "The Democrats' Real Turnout Problem," New York Time, November 17, 2016, Retrieved from https://www.nytimes.com/2016/11/20/opinion/sunday/the-democrats-real-turnout-problem.html?_r=2.
Link, M.W., M.P. Battaglia, M.R. Frankel, L. Osborn and A.H. Mokdad (2008), "A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) for General Population Surveys," *Public Opinion Quarterly*, 72(1): 6-27.
McCormick, J. (2016), "Clinton 3 Points Ahead of trump in Final Bloomberg National Poll," Bloomberg.com, November 7, 2016, Retrieved from https://www.bloomberg.com/politics/articles/2016-11-07/national-poll.
McDonald, M.P. (2007), "The True Electorate: A Cross-Validation of Voter Registration Files and Election Survey Demographics," *Public Opinion Quarterly*, 71(4): 588-602.
Miller, J. M., and J.A. Krosnick (1998). "The Impact of Candidate Name Order on Election Outcomes," *Public Opinion Quarterly*, 62, 291-330.
Mitofsky, W. (1998), "Review: Was 1996 a Worse Year for Polls than 1948?" *Public Opinion Quarterly*, 62 (2): 230-249.
Morning Consult (2016), "How We Constructed Our 50-State Snapshot," Retrieved from https://morningconsult.com/2016/09/08/constructed-50-state-snapshot/.
Mosteller, F., H. Hyman, P. McCarthy, E. Marks, and D. Truman (1949), "The Pre-election Polls of 1948: Report to the Committee on Analysis of Pre-election Polls and Forecasts," New York: Social Science Research Council.
Narea, A. (2016), "After 2016, Can We Ever Trust the Polls Again?" New Republic, December 14, 2016, Retrieved from https://newrepublic.com/article/139158/2016-can-ever-trust-polls-again.
Newport, F. (2016), "National Polling Accurately Nails Popular Vote," Gallup.com, November 23, 2016, http://www.gallup.com/opinion/polling-matters/198155/national-polling-accurately-nails-popular-vote.aspx.
Pasek, J., D. Schneider, J.A. Krosnick, A. Tahk, E. Ophir, and C. Milligan (2014), "Prevalence and Moderators of the Candidate Name-Order Effect: Evidence from Statewide General Elections in California," *Public Opinion Quarterly*, 78 (2): 416-439.

Pasek, J. (2016), "Another Reason Clinton Lost Michigan: Trump Was Listed First on the Ballot," University of Michigan Center for Political Studies Blog, November 29, 2016. Retrieved from http://cpsblog.isr.umich.edu/?p=1792.

Perry, P. (1973), "A Comparison of the Voting Preferences of Likely Voters and Likely Nonvoters," *Public Opinion Quarterly,* 37 (1): 99-109.

Pew Research Center (2009), "Understanding Likely Voters" January 22, 2009, Retrieved from http://www.people-press.org/files/2011/01/UnderstandingLikelyVoters.pdf.

Pew Research Center (2012), "Assessing the Representativeness of Public Opinion Surveys," May 15, 2012. Retrieved from http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/.

Pew Research Center (2016), "Flashpoints in Polling." April 1, 2016. Retrieved from http://www.pewresearch.org/2016/08/01/flashpoints-in-polling/.

Pew Research Center (2017), "Are Telephone Polls Understating Support for Trump?" March 31, 2017, Retrieved from http://www.pewresearch.org/2017/03/31/are-telephone-polls-understating-support-for-trump/.

Rothschild, D. and S. Goel (2016), "When You Hear the Margin of Error Is Plus or Minus 3 Percent, Think 7 Instead," New York Times, October 5, 2016. Retrieved from https://www.nytimes.com/2016/10/06/upshot/when-you-hear-the-margin-of-error-is-plus-or-minus-3-percent-think-7-instead.html.

Salvanto, A. (2016), "Poll: Women Propel Hillary Clinton into Battleground Lead Over Donald Trump." Retrieved from http://www.cbsnews.com/news/poll-women-propel-hillary-clinton-into-battleground-lead-over-donald-trump/.

Shashkevich, A. (2016), "Stanford Experts Discuss Polling Challenges During the 2016 Presidential Election Cycle," November 18, 2016, Retrieved from http://news.stanford.edu/2016/11/18/polling-challenges-presidential-election-cycle/.

Shepard, S. (2016), "How Could the Polling Be So Wrong?" Politico, November 9, 2016. Retrieved from http://www.politico.com/story/2016/11/how-could-polling-be-so-wrong-2016-231092.

Silver, N. (2014), "Registered Voter Polls Will (Usually) Overrate Democrats," FiveThirtyEight.com, September 9, 2014. Retrieved from https://fivethirtyeight.com/features/registered-voter-polls-will-usually-overrate-democrats/.

Silver, N. (2016a), "Why FiveThirtyEight Gave Trump A Better Chance Than Almost Anyone Else." Fivethityeight.com, November 11, 2016.

Silver, N. (2016b), "Pollsters Probably Didn't Talk To Enough White Voters Without College Degrees" FiveThirtyEight.com, December 1, 2016, Retrieved from https://fivethirtyeight.com/features/pollsters-probably-didnt-talk-to-enough-white-voters-without-college-degrees/.

Silver, N. (2017), "Why Early Voting Was Overhyped." FiveThirtEight.com, January 26, 2017, Retrieved from https://fivethirtyeight.com/features/early-voting-was-a-misleading-indicator/.

Smith, A. (2016), "UNH 2016 Election Polls." Presented at the New England AAPOR Chapter Election Postmortem held November 16, 2016.

Stinson, J. (2016), "The Pollster Who Foretold the Trump Tsunami," Lifezette.com, November 12, 2016, Retrieved from http://www.lifezette.com/polizette/pollster-foretold-trump-tsunami/.

Sturgis, P., N. Baker, M. Callegaro, S. Fisher, J. Green, W. Jennings, J. Kuha, B. Lauderdale, and P. Smith (2016), "Report of the Inquiry into the 2015 British General Election Opinion Polls," London: Market Research Society and British Polling Council.

Tedeschi, E. (2016), "The LA Times/USC Poll, Reweighted." November 8, 2016, Retrieved from https://sites.google.com/site/latuscrw/.

Trende, S. (2016), "It Wasn't the Polls That Missed, It Was the Pundits," RealClearPolitics.com, Retrieved from http://www.realclearpolitics.com/articles/2016/11/12/it_wasnt_the_polls_that_missed_it_was_the_pundits_132333.html.

Traugott, M.W. (2001), "Trends: Assessing Poll Performance in the 2000 Campaign," *Public Opinion Quarterly*, 65 (3): 389-419.

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

U.S. Census Bureau (2015), "2014 American Community Survey Research and Evaluation Report Memorandum Series #ACS 14-RER-30," January 8, 2015, Retrieved from https://www.census.gov/content/dam/Census/library/working-papers/2014/acs/2014_Walker_02.pdf.

Zukin, C. (2015), "A Primer on Election Polls: Or Why Different Election Polls Sometimes Have Different Results," Retrieved from https://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics/Election-Polling-AAPOR-2015-primary_cz120215-FINAL.aspx.

Wasserman, D. (2017), "2016 Popular Vote Tracker," The Cook Political Report, January 3, 2017, Retrieved from http://cookpolitical.com/story/10174.

Yourish, K. (2016), "Clinton and Trump Have Terrible Approval Ratings. Does It Matter?" New York Times, June 3, 2016. Retrieved from https://www.nytimes.com/interactive/2016/06/03/us/elections/trump-and-clinton-favorability.html?_r=0.

APPENDIX

Table A.0 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/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.

Table A.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	423	5.1	3.0
All battleground state polls	207	3.6	2.3
All non-battleground polls	206	6.4	3.3
Wisconsin	13	6.5	6.5
Ohio	13	5.2	5.2
Minnesota	5	4.9	4.9
Pennsylvania	24	4.2	4.2
North Carolina	18	4.8	4.0
Michigan	17	3.8	3.5
New Hampshire	16	5.0	3.4
Florida	23	2.9	1.3
Arizona	18	2.5	1.0
Georgia	14	2.3	0.9
Virginia	14	1.9	-0.2
Colorado	16	2.3	-1.6
Nevada	15	2.5	-1.7

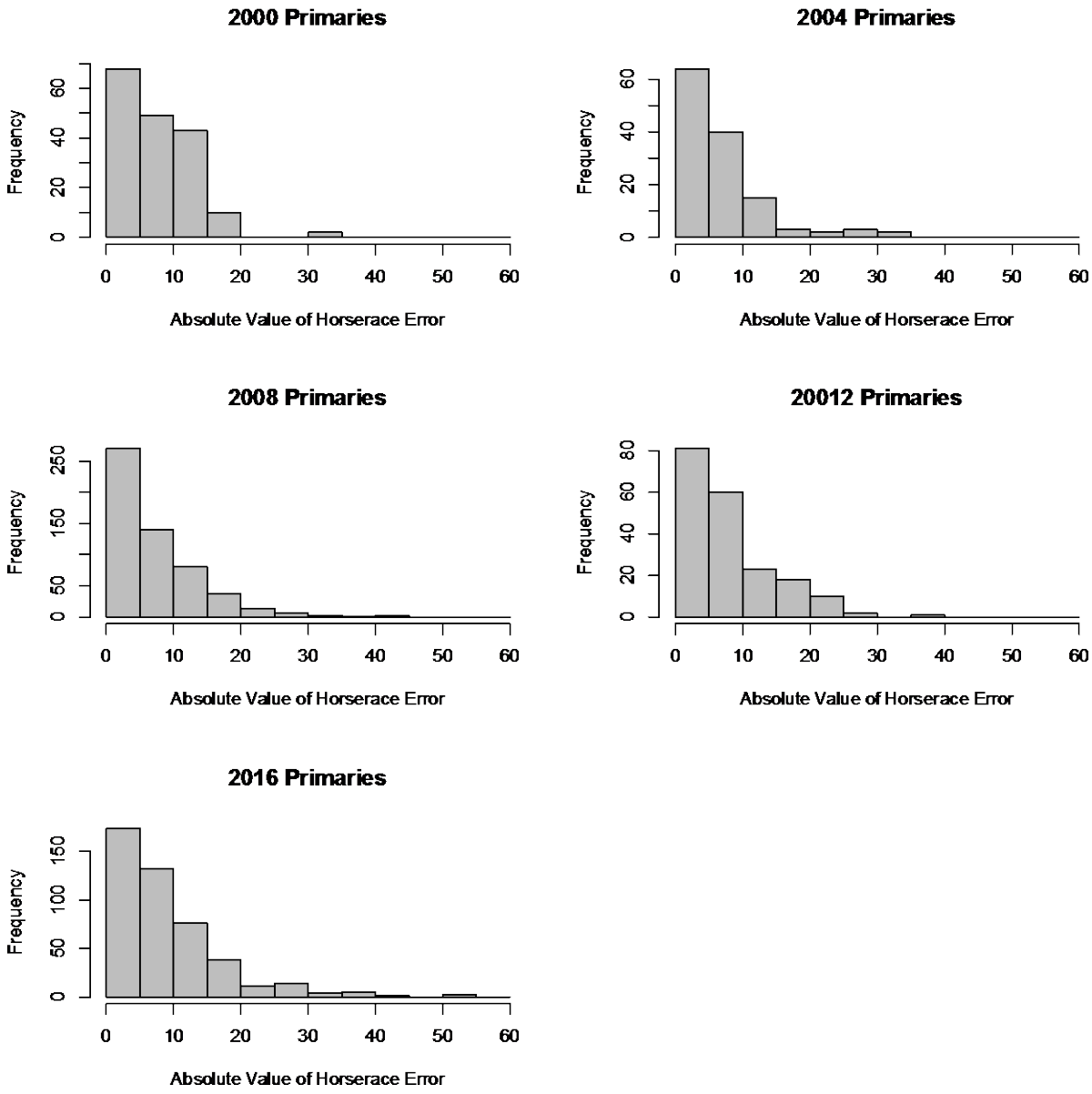


Figure A.1 Distribution of the Size of the Absolute Error in Primary Polls, 2000-2016. Note – Each bin is 5 percentage points wide.

A.A Poll Accuracy and the Margin of Error

The amount of error found in polls is important because it helps emphasize the importance of accurately accounting for, and reporting, the errors involved in polling. Most polls are accompanied by a "margin of error" that is often interpreted as the amount of error in the poll.

This is an incorrect interpretation. The margin of error denotes how much error is likely due to sampling variation alone such that if the survey were to be redone 100 times under exactly the same conditions, 95 of the estimates would lie within that range. The typical margin of error understates even this assessment because it reflects the amount of variation we expect in estimating a single proportion by using a sample of respondents from a larger (fixed) population. In a horserace poll, however, the quantity of interest is often the difference in support for two candidates. The error due to sampling variability alone when estimating the difference in two (correlated) quantities from the same poll is therefore larger than what the margin of error reports.

The amount of sampling variability in a poll is not the same as a description of how close the poll's prediction may be to the truth. It is wrong to equate the margin of error with the amount of polling error. There are many other sources of error that can affect the accuracy of a poll and whose effects are not reflected in the margin of error. For pre-election polls, these difficulties typically involve issues such as the possibility of systematic non-response, due to either technological (e.g., access to the Internet or a land line telephone) or psychological (e.g., distrust of the media and pollsters leading to a refusal to participate) reasons, as well as the additional difficulty of identifying "likely voters" and what the composition of the electorate will be.

For example, 2016 primary polls had an average of 636 respondents, which yields a margin error of ±4% using standard assumptions and calculations (which do not account for the loss in precision due to weighting or design departures from a simple random sample of the population). However, the average absolute error in the margin of victory was 9.3% – more than twice the stated margin of error. The fact that the average error was so much greater than the margin of error highlights the importance of better understanding and communicating exactly what the margin of error is and is not. It is not a statement about the potential error that the poll contains and conflating these concepts does a disservice to our ability to interpret and assess the accuracy of pre-election polls.

Table A.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).

A.B Regression Analysis Examining Effects of Poll Design Features on Accuracy

The focus of our evaluation 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 A.2 presents the performance of polls within each primary contest.

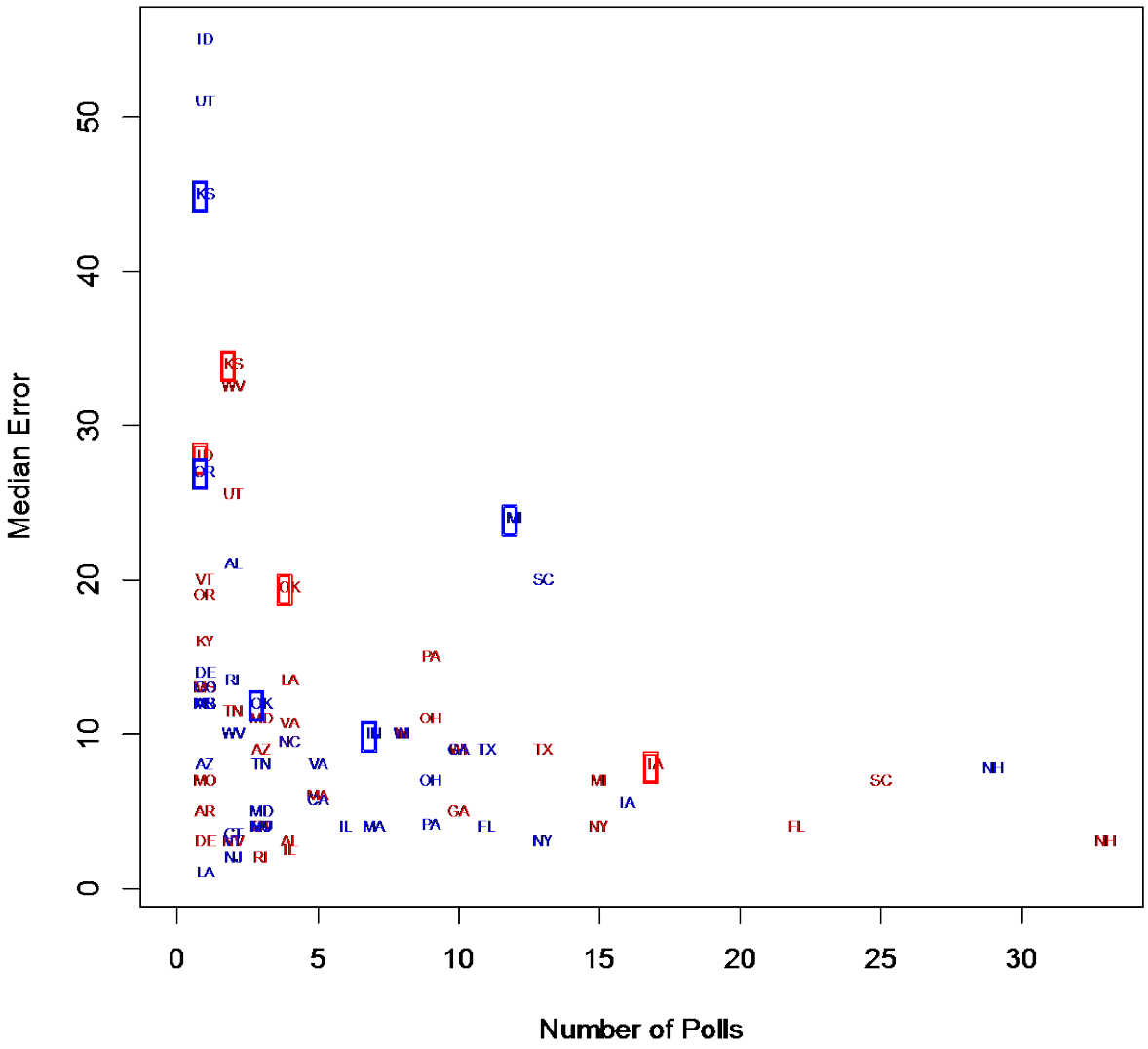


Figure A.2 Median Absolute Error in Primary Contests: Note – Republican (red) and Democrat (blue) results for each state are plotted.

Each labeled point in Figure A.2 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 A.2. 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 A.3 presents the distribution of polling errors by mode.

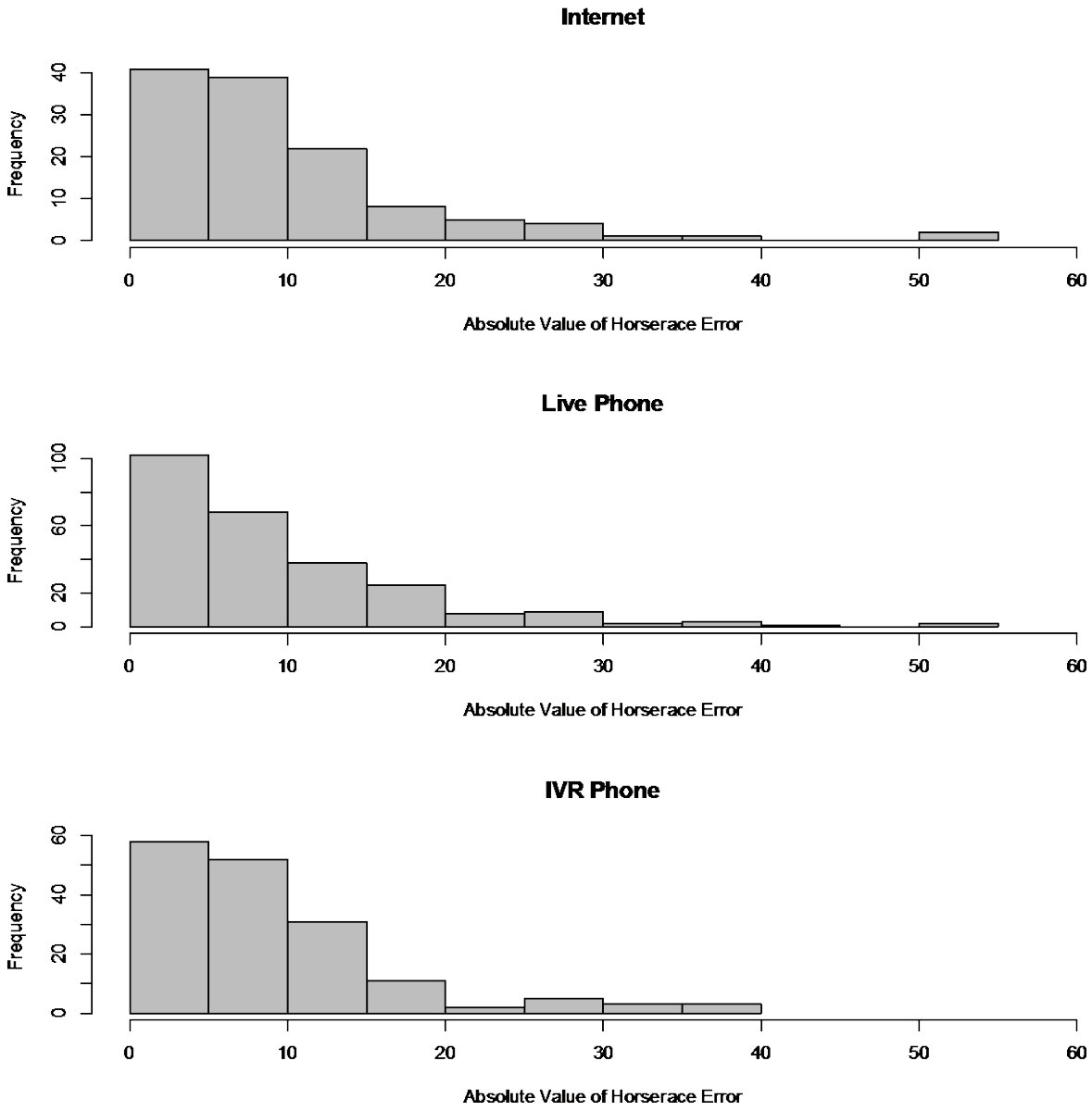


Figure A.3 Absolute Error by Mode

Figure A.3 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 A.4 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.

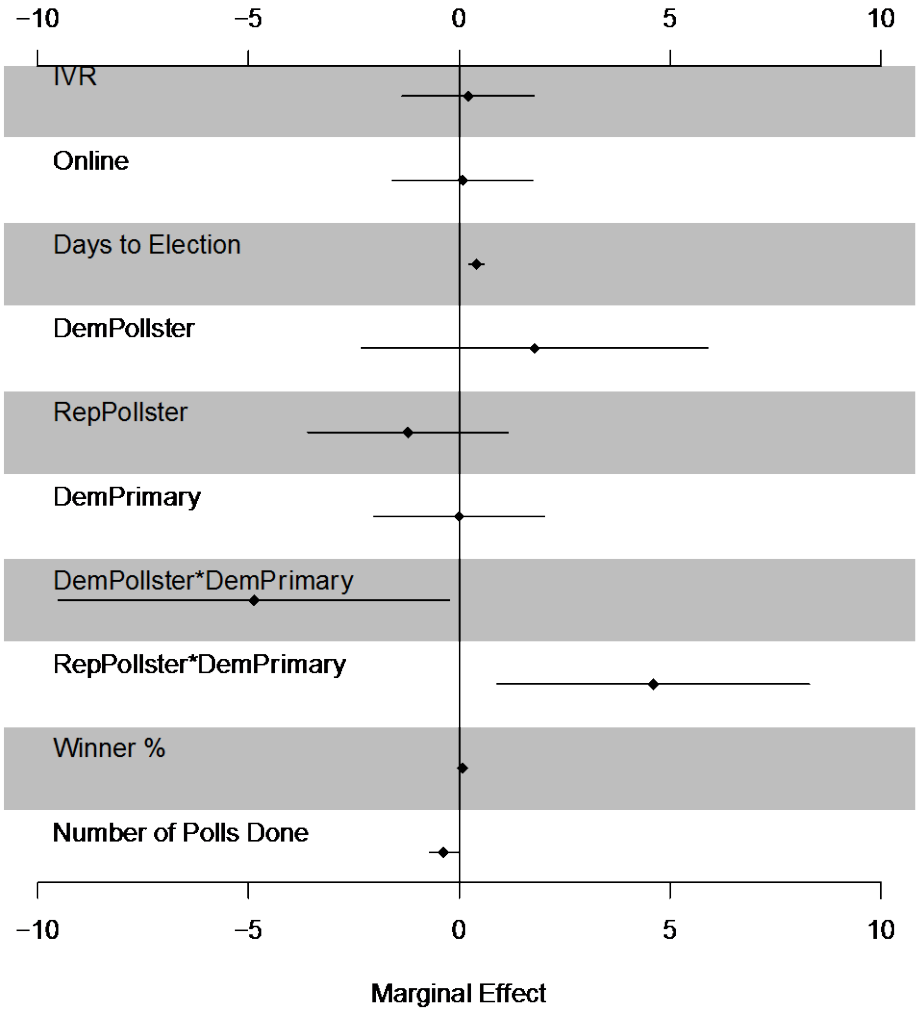


Figure A.4 Marginal Effect of a One-Unit Change in Each Feature on the Absolute Error for 2016 Primary Polls

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.[19] 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. 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 A.1.

[19] 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.

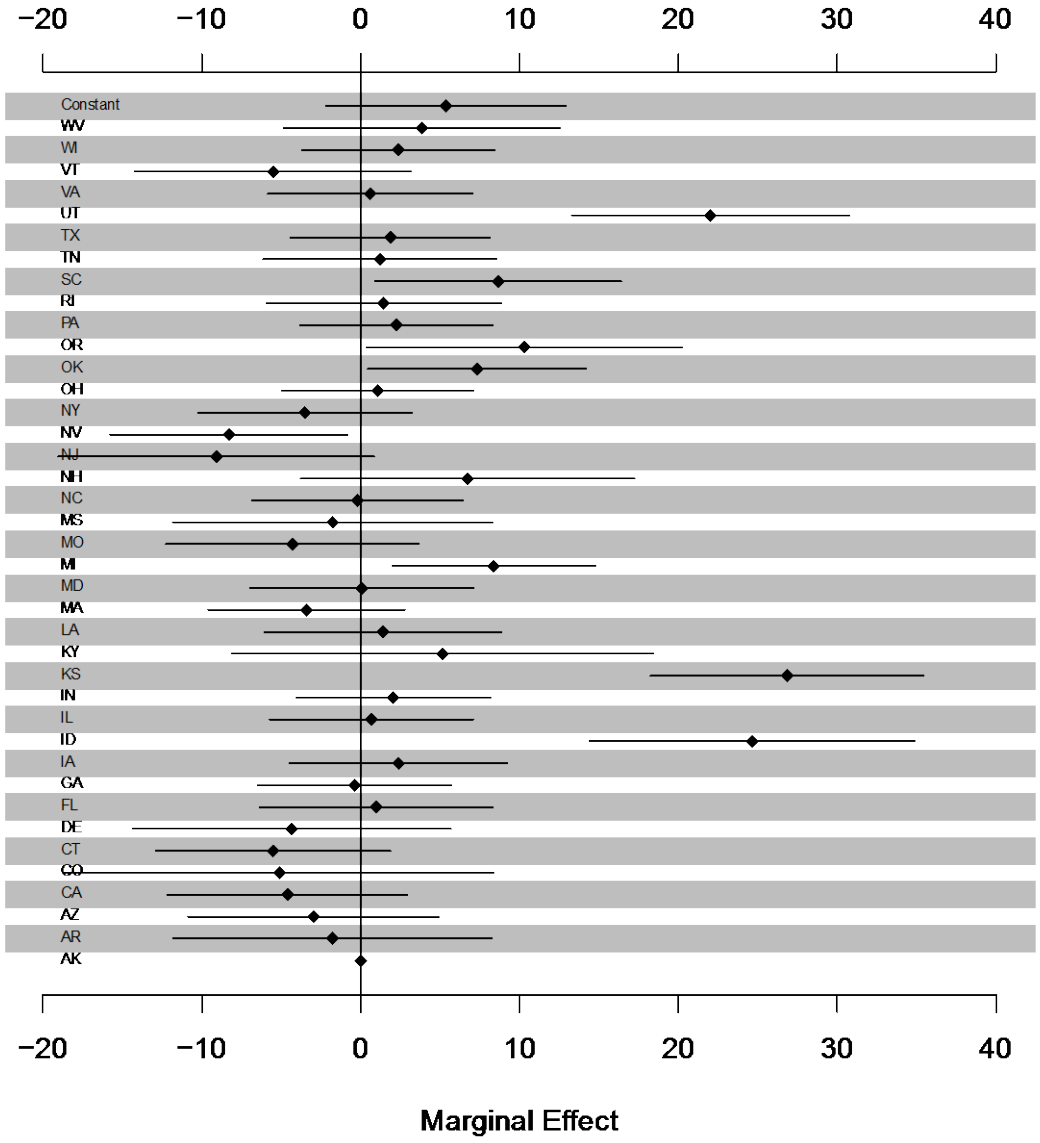


Figure A.5 Average Error in Polls After Controlling for Poll Characteristics

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 A.3 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.[20] Caucuses are associated with much bigger errors – the average absolute horserace is nearly 10 points greater in caucuses and this is a statistically 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.

[20] Mixed primaries are the omitted category.

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.

A.C 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.

A.D 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.[21] 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, 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:

$$(1^{st} \text{ place vote} - 2^{nd} \text{ place vote}) - (1^{st} \text{ place poll} - 2^{nd} \text{ place poll}).$$

[21] 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.

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 A.4 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 A.4 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 A.4 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.[22] Table A.5 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 A.5 Selected Correlates of Primary Poll Performance

	Winner's vote-poll margin	Winner's poll margin

[22] 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.

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)	–
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 A.6 and A.7. 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.

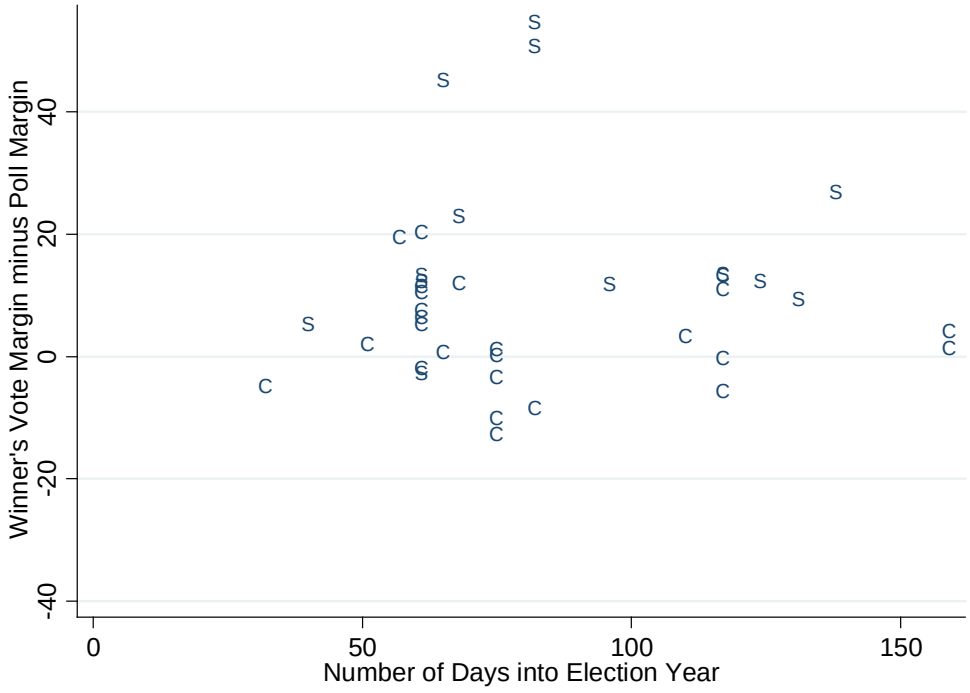


Figure A.6 The Polls and the Vote in the 2016 Democratic Presidential Primaries. Notes – Each entry in the figure is the difference in a state between the winner's actual vote share and the share of the second place candidate minus the corresponding pre-election poll margin in the two weeks leading up to the election. A "C" indicates a Clinton win and an "S" represents a win by Sanders.

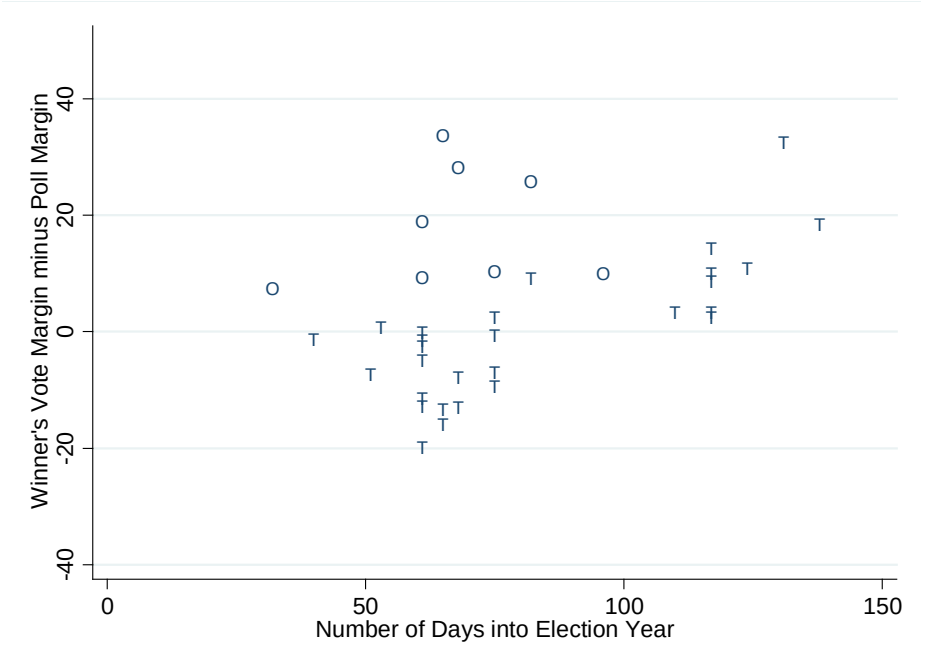


Figure A.7 The Polls and the Vote in the 2016 Republican Presidential Primaries.
Notes – Each entry in the figure is the difference in a state between the winner's actual vote share and the share of the second place candidate minus the corresponding pre-election poll margin in the two weeks leading up 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 A.6. The first column contains results of a simple baseline regression containing the winner's poll margin. As expected given Table A.5, 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 A.6 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 A.6 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 A.7 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 A.7. 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.[23]

Table A.7 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

[23] 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.

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 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.

A.E Testing for *Shy Trump* Mode Effects in National Polls Conducted September 1st and Later

The main 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 A.8. 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 A.8, 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.[24] Finally, more than 90 percent of the polls used likely voter, and there was no difference between modes on this factor.[25] Table A.9 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 A.9 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

[24] 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.

[25] 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.

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[26]. 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.

Table A.9 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

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

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. Methodological characteristics explain 10%-13% of the variance is estimates.

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.

Another possibility is that Trump supporters were more likely than other respondents either to report being undecided or to refuse to reveal their preference. In this case, there should see be a relationship between the proportion of nondisclosers (all those who do not reveal their preference) and the proportion of Trump supporters in the polls and no such relationship for Clinton. However, we need to be careful because the proportion of nondisclosers is related to the methodological characteristics of the polls. First, as shown in table 1, there is a significant relationship between the proportion of nondisclosers in the polls and mode of administration. The average proportion of nondisclosers is 6.6% on average but it is 4.3% for polls using live phone, 5.6% for IVR/online polls and 8.5% for Web polls.

Table A.10 Determinants of the Proportion of Nondisclosers

Intercept	5.33	***
Time	0.01	
Time squared	0.00	
Time cubic	0.00	
Days in field	-0.37	**
Use of LV model	0.78	

Tracing poll	1.53	**
Live phone	0.08	
Online poll	3.67	***
Explained variance	*41.5%*	

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

A regression with proportion of nondisclosers as the dependent variable, presented in Table A10, shows that there is no significant change in the proportion of nondisclosers over time. Two characteristics of the polls appear related to the proportion of nondisclosers. Polls that stay in the field a larger number of days tend to show a lower proportion of nondisclosers (-.75 point per day in the field) while Web polls show proportions that are on average more than 3.6 points higher than IVR/online polls. These features explain a total of 41.5% of the variance in the proportion of nondisclosers.

If we control for all these features, what is the relationship between proportion of nondisclosers in the polls and estimates of support for the different candidates? The analysis shows that the proportion of nondisclosers is not related to the estimates of the support for Trump compared to Clinton. However, if we consider the estimates for all the candidates, we see that the higher the proportion of nondisclosers in a poll, the higher the estimates of support for Trump and for Clinton and the lower the support for the third party candidates. The proportion of nondisclosers explains more than 8% of the variance in estimates of the third party candidates.

Live phone polls use sampling frames that overlap. They combine lists of landline and cell phone telephone numbers. Since some of the cell phone users may also be joined by landline phone, they are more likely to be interviewed. Although this overlap may be corrected by weighting, this procedure may not totally compensate. Do people who can be joined both by landline and cell have different characteristics that are related to vote intention? If this is the case, it could explain the fact that support for Clinton is higher in live phone polls than in IVR/online polls where there is no such overlap of sampling frames.

If this hypothesis is true, there should be a relationship between the proportion of cell phones in the samples and estimates of support. More than a third of the polls (n=58) use the live phone mode and 57 provided the information on the proportion of cell phones in their samples. The average proportion of cell phones in these polls was 57% with a standard deviation of 10.6. The lowest proportion was 25% and the highest, 75%. The most common proportion was 65% (22% of the polls). *There is no consensus regarding the proportion of cell phones that should be included in the samples*.

Table A.11 and Table A.12 show the results of the analyses of support for the candidates, controlling for change over time and for other methodological features of the polls. It shows that there is no impact of the proportion of cell phones on estimates of any of the candidates, whatever the comparison used. The variables that are significant – essentially the time variables and the fact that the poll is a tracking poll – explain between 26% and 37% of the variation in estimates of support for Trump and around 10%-18% of the variation in support for Clinton and for the other candidates. Therefore, the overlap in the sampling frames cannot explain the

differences between the live phone polls and the IVR/online polls. It, however, questions the impact of the inclusion of cell phones in the samples.

With all the polls conducted online in the same category, we have seen that, on average, web polls tend to trace a different portrait of change over time in support for Trump over the two main candidates. Tables A.11 and A.12 confirm that the estimates of change over time differ between live phone and web polls. Estimates of the linear and quadratic trends are less than half for Web polls than for live phone polls.

Table A.11 Support for Trump from September 1st to Election Day

	On the sum of the two main candidates				On the sum of all candidates			
	Live phone only		Web only		Live phone only		Web only	
Intercept	46.49	**	45.52		40.08		38.01	
Time variables		*						
Time	-0.11	**	-0.06		-0.12	*	-0.05	
Time squared		*		**		*		*
Time cubic	0.00	*	0.00	*	0.00	*	0.00	
Explained variance	0.00	*	0.00	*	0.00	*	0.00	**
	25.1%		*15.4%*		*23.6%*		*4.7%*	
Methods variable								
Days in field	-0.17		0.16		-0.23		0.42	**
Used LV model								*
Tracking poll	-0.50		-0.59		0.18		-0.68	
Explained variance	2.56	**	1.00		1.50		2.70	**
		*						*
	37.9%		*15.3%*		*26.3%*		*25.5%*	
Variables specific to modes								
% cell (live phone)	0.02				0.06			
Panel (online)			1.50	**			2.71	**
River sampling				*				*
Explained variance			-0.16				0.48	
	37.6%		*28.4%*		*30.0%*		*53.5%*	
N	58		80		57		80	

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

Table A.12 Support for Clinton and Other Candidates from September 1st to Election Day

	Clinton				Other candidates			
	Live phone only		Web only		Live phone only		Web only	
Intercept	45.68	***	45.98	***	14.24	***	6.01	***
Time variables								
Time	0.08		0.06	**	0.04		-0.01	
Time squared	0.08		0.00	**	0.00		0.00	
Time cubic	0.00		0.00		0.00		0.00	***
Explained variance	9.4%		28.3%		3.3%		13.4%	
Methods variable								
Days in field	0.06		0.15		0.17		-0.58	***
Used LV model	1.30		0.30		-1.48		0.38	
Tracking poll	-3.17	***	0.93+		1.66		-3.63	***
Explained variance	17.8%		47.5%		3.3%		54.7%	
Variables specific to modes								
% cell (live phone)	0.03				-0.09			
Panel (online)			0.05				-2.76	***
River sampling			0.91				-1.39	*
Explained variance	17.2%		48.3%		4.4%		75.1%	
N	57		80		57		80	

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

\+ Highly significant before entering panel and river sampling (14% of variance, b=1.67)

However, the general category of online polls is quite heterogeneous. Our hypothesis was that since most online polls use panels, they may have pools of respondents that are more homogenous and that this may be related to estimates that differ from those produced by other modes of administration. Two characteristics of online polls are related to homogeneity. Most online polls use panels of respondents while others sample pools of respondents that differ with each poll. In addition, some pollsters complement their samples with river sampling[27]. Both these features are indicators of heterogeneity. The fact that panel recruitment is probabilistic instead of opt-in could also play a role but only one pollster conducting two polls during the period used a probabilistic panel. It is therefore impossible to test this possibility. Finally, the number of panel members could also be related to homogeneity but we could not get this information for all the pollsters.

[27] River sampling is a way of recruiting respondents that are not in the original samples using a procedure that ask internet users selected at random.

Table A.11 shows that, all else being equal, using a panel to conduct online polls leads to estimates of support for Trump that are 1.5 points higher if we consider the sum of the two main candidates and 2.7 points higher when we consider all the candidates. The results of the two analyses differ. In the first case, specific methodology explains by itself 13% (28.4-15.3) of the variance and only the use of a panel is significant. All the other characteristics of the polls are not significant. In the latter case, estimates of support for Trump appear to be half a point higher per day in the field and 2.7 points higher in tracking than in non-tracking polls. Specific methodological factors explain 28% of the variance.

When we examine the estimates of support for Clinton and for the third party candidates presented in Table A.12, we conclude that the use of a panel and of river sampling are not related to estimates of support for Clinton. They are only associated with estimates of support for the third party candidates. The coefficients show that the estimates of support for the third party candidates are more than half a point lower per day in the field, 3.6 points lower in tracking polls. They are 2.8 points lower in surveys using panels and 1.4 points lower when river sampling is used. Specific methodological factors explain close to 20% of the variance.

Therefore, we may conclude that indicators of homogeneity are only related to the estimates of the relative share of support for Trump compared with the third party candidates. More possible homogeneity is associated to higher support for Trump and lower support for the third party candidates, an estimate that is closer to the final vote. It may be that web polls using panels have more control over their samples or their weighting/adjustment procedure.

A.F Polling Aggregators in the Primaries

We examined three polling aggregators, resulting in information from five different estimation methods.

Aggregator methods

First, FiveThirtyEight had three different approaches to calculating predictions. For races where limited polling information was available, the FiveThirtyEight Polling Average is a simple weighted average of the available polls. For example, in the Missouri Republican Primary, only one poll was conducted during 2016 (March 3-10) and used in the FiveThirtyEight polling average (two other polls were conducted in 2015 – one in December and one in August).[28] The FiveThirtyEight Polls Only predictions are based only on data that come from the polls themselves; FiveThirtyEight Polls Plus incorporates additional information into the prediction, including information from national polls and endorsements. We focus primarily on the FiveThirtyEight Polls Only and FiveThirtyEight Polls Plus predictions. FiveThirtyEight includes all polls unless it was conducted by a campaign, an affiliated PAC or super PAC, or is on the list of FiveThirtyEight banned pollsters. If a poll publishes estimates based on multiple populations (likely voters, registered voters, all adults), FiveThirtyEight limits the poll estimates to the closest representation of likely voters reported by the poll.

Second, Huffington Post Pollster uses information from the polls and Cook Political Report to develop its predictions. To be included in the Pollster estimates, the poll has to meet a set of

[28] http://projects.fivethirtyeight.com/election-2016/primary-forecast/missouri-republican/

methodological disclosure requirements from the National Council on Public Polling, including dates of study, information about sponsor, field dates, mode, sample frame, sample size, and question wording, among other information. Pollster excludes landline-only polls as well as polls that do not meet an editorial evaluation of adequate quality. Furthermore, HuffPollster excludes polls that do not ask about the horserace with closed-ended questions, and uses only unique sample points (excludes overlapping samples in rolling averages). If a poll publishes estimates based on multiple populations (likely voters, registered voters, all adults), HuffPollster also limits the poll estimates to the closest representation of likely voters reported by the poll. Finally, RealClearPolitics uses a simple unweighted average of polls. RealClearPolitics does not have a clear statement about which polls are considered eligible to include or exclude from its estimates.

Aggregator Errors

Overall, the aggregators underestimated the size of the margin between the top two candidates. As shown in Table A.13, looking at all races, the average signed error in the horserace margin was -4.65, indicating that the predictions underestimated the margin by 4.65 percentage points. When the analysis is restricted to the contests where four estimates were made, the signed error decreases slightly to -3.99, so that the aggregators underestimated the margin by 3.99 percentage points. There were no significant differences between the aggregators in their signed prediction accuracy (F=0.04, p=0.99).

Turning to the absolute error, for all races, the average error across all of the aggregators was 8.32, indicating that the average difference between the margin calculated by the aggregators and the actual margin for the winner was 8.32 percentage points. When the analysis is restricted only to the contests where four estimates were made, the signed error drops by one percentage point to 7.34. Although significant differences appear across the aggregators in the absolute horserace error overall (F=5.92, p=0.0002), this is explained by different aggregators making predictions in different races. When only common races are examined, there are no significant differences across the aggregators in the average absolute error (F=0.09, p=0.96)

Table A.13 Mean Signed Horserace Error Overall and by Aggregator, All Contests and Contests with Predictions from All Firms

	Signed Horserace Error					Absolute Horserace Error				
	N	Mean	Std. Dev.	Min.	Max	N	Mean	Std. Dev.	Min.	Max
All contests										
Overall	230	-4.65	10.76	-61	18	230	8.32	8.25	0	61
FiveThirtyEight Polling Average	4	-18.75	31.91	-52	11	4	27.25	22.29	6	52
FiveThirtyEight Polls Only	59	-4.64	9.60	-35	12	59	7.90	7.12	0	35
FiveThirtyEight Polls Plus	59	-4.23	9.77	-41	11	59	7.66	7.35	0	41
Huffington Post	53	-3.77	11.99	-61	18	53	8.57	9.13	1	61
RealClearPolitics	55	-4.93	8.88	-32	12	55	7.87	6.35	0	32
F Test	F(4,225)=1.32, p=0.27					F(4,225)=5.92, p=0.0002				

Common contests										
Overall	183	-3.99	8.69	-41	18	183	7.34	6.11	0	41
FiveThirtyEight Polls Only	46	-4.13	8.84	-35	12	46	7.26	6.45	1	35
FiveThirtyEight Polls Plus	46	-3.98	9.48	-41	11	46	7.33	7.15	0	41
Huffington Post	45	-3.64	8.89	-23	18	45	7.73	5.61	1	23
RealClearPolitics	46	-4.20	7.77	-23	12	46	7.07	5.23	0	23
F Test	F(3,179)=0.04, p=0.99					F(3,179)=0.09, p=0.96				

Note: The Huffington Post did not make predictions for the Republican contest in California for candidates other than Trump on the final prediction date.

Error by Number of Aggregators Predictions. In general, contests for which more of the aggregators made predictions had smaller signed horserace errors than contests in which only one or two aggregators made predictions (Table A.14). The contests for which only one or two predictions were made included the Republican contests in Alaska (March 1), Alabama (March 1), Kansas (March 5), Kentucky (March 5), Missouri (March 15), New Jersey (June 7), Tennessee (March 1) and West Virginia (May 10) and the Democratic contests in Alabama (March 1), Missouri (March 15), and Utah (March 22).

Table A.14 Mean Signed and Absolute Horserace Error Overall and by Contest, by Number of Aggregators with Predictions

Number of	Overall			Republican			Democrat		
predictions	N	Mean	SD	N	Mean	SD	N	Mean	SD
Signed Error									
1	4	-9.25	30.02	3	5.00	11.53	1	-52.00	–
2	13	-14.00	21.94	9	-18.00	25.43	4	-5.00	6.68
3	30	-4.03	10.11	9	-6.11	12.33	21	-3.14	9.20
4	183	-3.99	8.70	91	-3.49	9.31	92	-4.48	8.06
		F=3.92	P=0.009		F=5.04	P=0.003		F=11.25	P<.0001
Absolute Error									
1	4	20.25	21.34	3	9.67	5.51	1	52	–
2	13	18.15	18.35	9	23.78	19.37	4	5.50	6.14
3	30	8.43	6.75	9	8.78	10.36	21	8.29	4.79
4	183	7.34	6.11	91	7.36	6.65	92	7.33	5.57
		F=11.11	P<.0001		F=10.06	P<.0001		F=22.39	P<0.0001

Looking at all of the contests together, the average signed horserace error for races for which there was only one or two predictions were -9.25 and -14.00, respectively, compared to -4.03 and -3.99 for contests for which there were 3 or 4 predictions (F=3.92, p=0.009). When we look at the Republican contests only, 91 of the 112 contests (81.25%) had predictions from all four aggregators. These 91 contests had a mean signed horserace error of -3.49 and -6.11 for the 9 contests with three predictions, compared to larger values for Republican contests that had two or fewer predictions (F=5.04, p=0.003). There were 92 Democratic contests in which there were predictions from all four aggregators, making up 77.97% of the total. Here, the average horserace error was -4.48 for contests with four predictions and -3.14 for contests with three predictions, with larger values for the few other predictions.

We see a similar pattern when examining absolute horserace error for the for the number of predictions – contests for which there were more predictions had lower absolute error, on average, than contests for which there were fewer predictions. Overall, the absolute error for contests with only one prediction was 20.25 percentage points, falling slightly to 18.15 percentage points for contests with two predictions, 8.43 percentage points for contests with three predictions, and 7.34 for contests with 4 predictions ($F=11.11$, $p<.0001$). This pattern generally held for both the Republican ($F=10.06$, $p<.0001$) and Democratic ($F=22.39$, $p<.0001$) contests. The average absolute error for the Republican contests with four predictions was 7.36 percentage points, compared to 8.78 for contests with three predictions, 23.78 for contests with two predictions, and 9.67 for contests with only one prediction. On the Democratic side, the average absolute error was 7.33 for contests with four predictions, and increasing to 8.29, 5.50, and 52 for contests with three, two, and one predictions, respectively.

Error by Contest and State. There was no significant difference in the signed horserace error between the Democratic and Republican contests for the aggregators, either overall (Republican contests = -4.64; Democratic contests = -4.66; $t=-0.01$, $p=0.99$) or for common contests (Republican contests=-3.49; Democratic contests = -4.48, $t=-0.76$, $p=0.45$).

There were, however, significant differences in the signed horserace error across states. Figure A.8 shows the signed horserace error, averaged across aggregators, for each state for the Republican and Democratic contests by the number of predictions for each state. In most states, the average of the aggregator predictions understated the actual margin – where there were four predictions, in only six states did the average across the aggregators overstate the margin between the two candidates for the Republican race and in five states for the Democratic race.

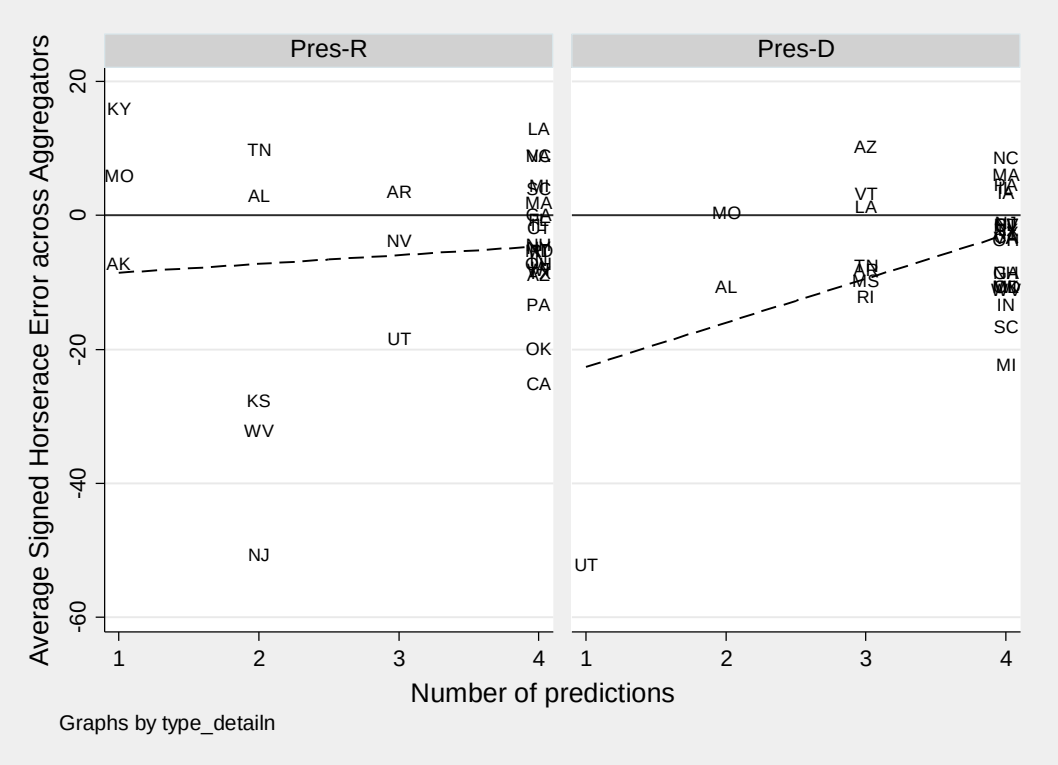


Figure A.8 Average Signed Error Across Aggregators, by Contest, State, and Number of Predictions.

Similarly, there were no significant differences in absolute error of the aggregators' predictions between the Democratic and Republican contests overall (Republican contests = 8.86, Democratic contests = 7.81, t=-0.96, p=0.34) or for the common contests (Republican contests = 7.36, Democratic contests = 7.33, t=-0.04, p=0.97).

As with signed error, there are significant differences in absolute error by state. Figure A.9 shows the signed absolute error for each state across the Republican and Democratic contests for states by the number of predictions for each state. On the Republican side, there were four states with four predictions in which the average absolute error from the prediction was at least 10 percentage points different from that of the actual margin – California (25.67), Louisiana (13), Oklahoma (19.75), and Pennsylvania (13.25). On the Democratic side, six states with four predictions exceeded a 10 percentage point difference between the average aggregator predictions and the actual margin – Indiana (13.25), Maryland (10.50), Michigan (22.25), Oklahoma (10.50), South Carolina (16.5), and Wisconsin (10.50).

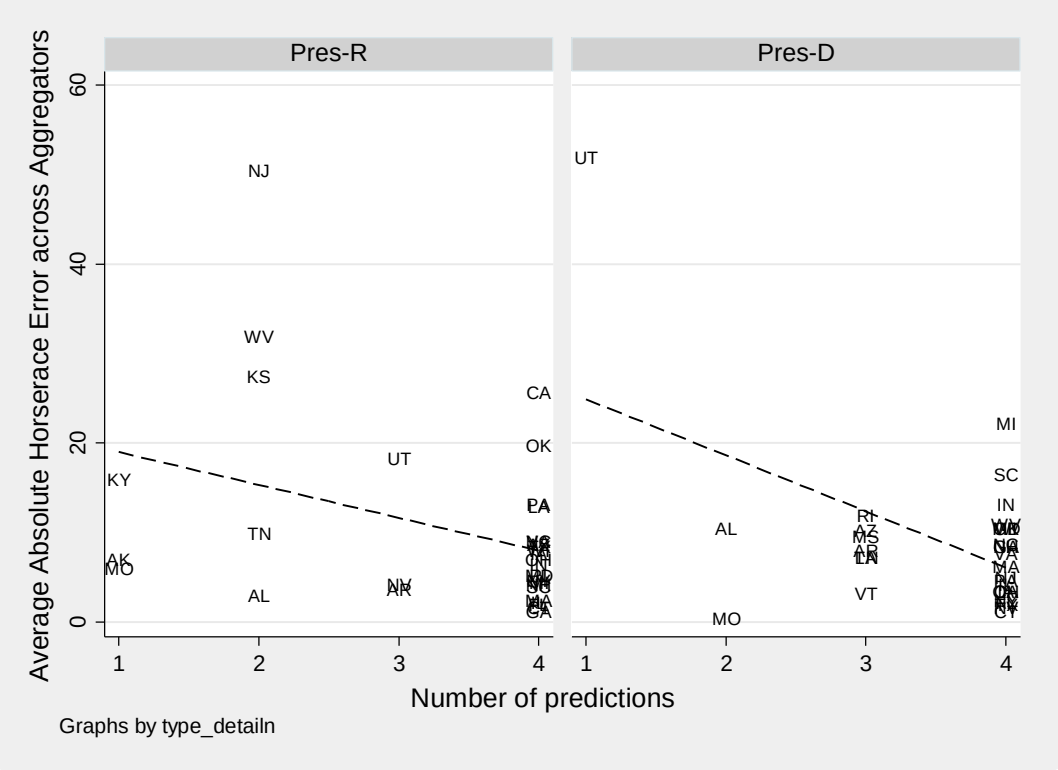


Figure A.9 Average Absolute Error Across Aggregators, by Contest, State and Number of Predictions

Error by winner's percentage. Looking only at the 46 state contests with four predictions, there is a negative association between the winner's percentage of the final vote and the signed horserace error. Overall, the correlation is -0.34 (p=0.02); that is, the larger the winner's percentage, the more that the prediction underestimated the difference between the percentage for the winner and the runner up. Examining the 23 Republican and 23 Democrat races

separately, the correlation is r=-0.45 (p=0.03) for the Republican races and a non-significant r=-0.22 (p=0.31) for the Democratic races (Figure A.10).

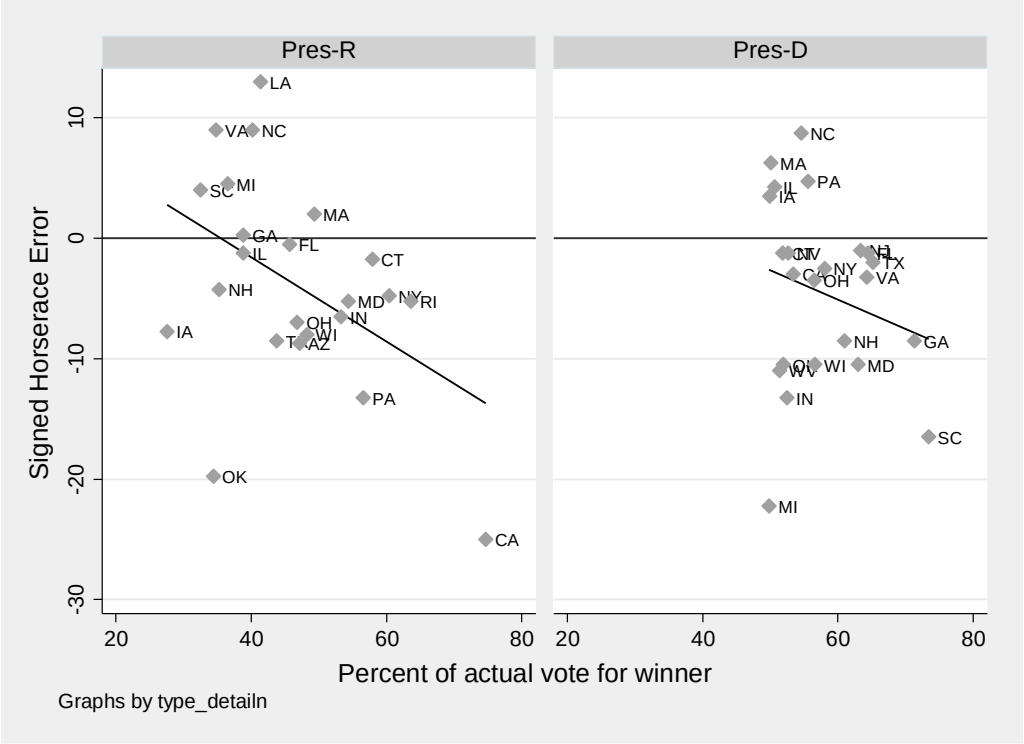


Figure A.10 Average Signed Error Across Aggregators by Percent of Actual Vote for Winner, by Contest

We see a slightly different story for the association between absolute error and the percent of the actual vote for the winner, as shown in Figure A.11. Overall, there is no association between absolute error and the percent of the actual vote for the winner (r=0.15, p=0.32). We also see no association between absolute error and the percent of the actual vote for the winner in the Republican (r=0.26, p=0.22) and Democratic races (r=0.06, p=0.79).

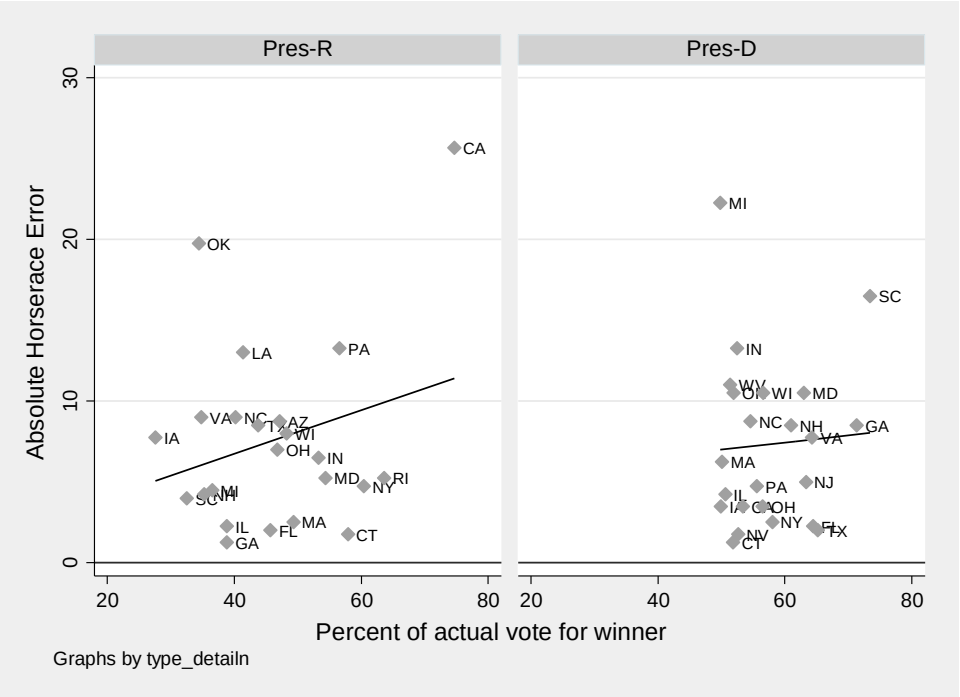


Figure A.11 Average Absolute Error Across Aggregators by Percent of Actual Vote for Winner, by Contest

Correct Projection. Across all 232 projections, 23 (9.91%) were called incorrectly. Incorrect calls were equally likely to happen on both the Republican and Democratic sides of the primaries (χ^2=0.02, p=0.90) - 9.65% of the projections on the Republican side (11 out of 114) were called for the wrong candidate and 10.17% (12 out of 118) were called for the wrong candidate. When looking only at states with four predictions, 18 of the 184 contests (9.78%) were called incorrectly. As with overall, incorrect calls were equally likely to occur for both Republican (8 of 92, or 8.70%; all of the calls in Iowa and Oklahoma) and Democratic (10 of 92, or 10.87%; all of the calls in Indiana, Michigan, and two of the calls in Oklahoma) contests. Each of the aggregators was equally likely to have an incorrect call overall (χ^2=0.65, p=0.96) and for commonly called contests (χ^2=0.25, p=0.97).

As shown in Table A.15, the average signed error of the aggregator predictions of the margin is significantly larger for contests that were called incorrectly (-15.74) than for contests that were called correctly (-3.42; t=5.53, p<.0001) overall and for only common contests (-15.22 vs. -2.76; t=6.37, p<.0001). This pattern and magnitude holds for both measures of error for both the Republican and Democratic contests.

Table A.15 Average Signed and Absolute Error Overall and by Contest, by Number of Aggregators with Predictions

	Signed Error			Absolute Error		
	Incorrect calls	Correct Calls	t	Incorrect calls	Correct Calls	t
All Contests						
Overall	-15.74	-3.42	5.53****	15.74	7.48	-4.76****
Republican	-15.64	-3.45	3.29***	15.64	8.12	-2.54*
Democratic	-15.83	-3.40	4.81****	15.83	6.91	-4.69****

Common contests						
Overall	-15.22	-2.76	6.37****	15.22	6.48	-6.35****
Republican	-13.75	-2.51	3.45***	13.75	6.75	-2.96**
Democratic	-16.40	-3.02	5.76****	16.40	6.22	-6.61****

Note: *p<.05, **p<.01, ***p<.001, ****p<.0001

Errors and white non-college population in each state. One hypothesis for why the poll estimates differed from the actual margin is differential nonresponse bias by systematically missing white voters without a college degree (Silver 2016b). If this is the case in the primary contests, then we should see that the errors in the difference in the margin are systematically associated with the percent of white noncollege voters in each state.

Overall, the correlation between the share of white noncollege voters in each state and the signed horserace error is small and not statistically different from zero (r=-0.05, p=0.66). When we look at the Republican and Democratic races separately, we see similar results. For the 34 Republican contests, there is a nonsignificant small positive association (r=0.06, p=0.75) between the share of white noncollege voters and the mean signed error across the aggregators. For the Democratic contests, there is a nonsignificant negative association (r=-0.20, p=0.28) between the share of white noncollege voters and the mean signed error across the aggregators.

When we instead examine the actual values for each of the 112 Republican predictions and the 118 Democratic predictions from each aggregator rather than the mean horserace prediction error across aggregators. Here, as visualized in the right panel of Figure A.12, the correlation between the signed horserace error and the percent of white noncollege voters for the Republican contests weakens (r=-0.007, p=0.94) and reaches statistical significance for the Democratic contests (r=-0.19, p=0.03). Because the signed errors start close to zero, this negative correlation indicates that the states with more white non-college voters were more likely to underestimate the margin for the Democratic side, and this pattern did not hold on the Republican side.

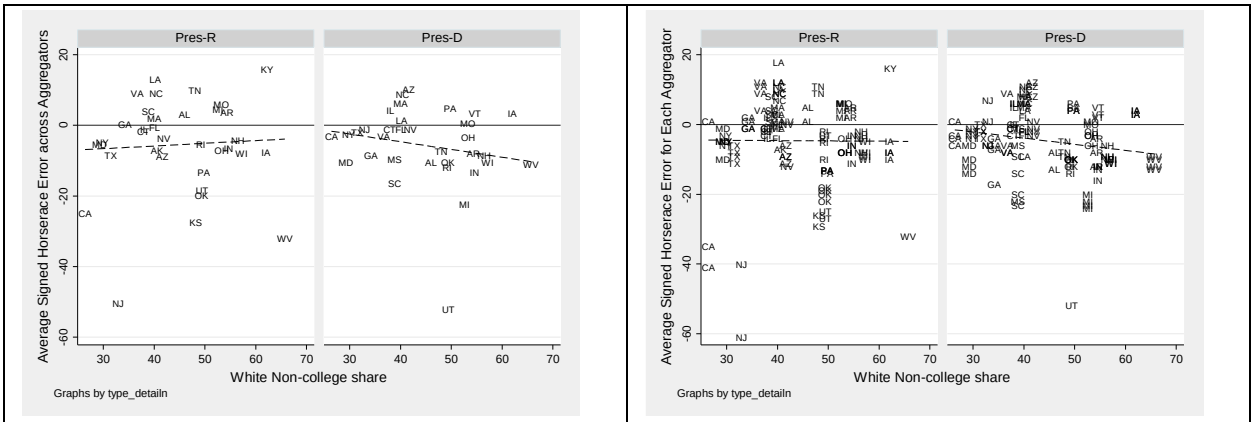
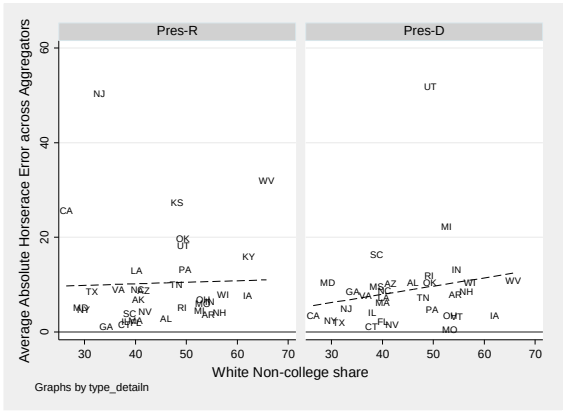


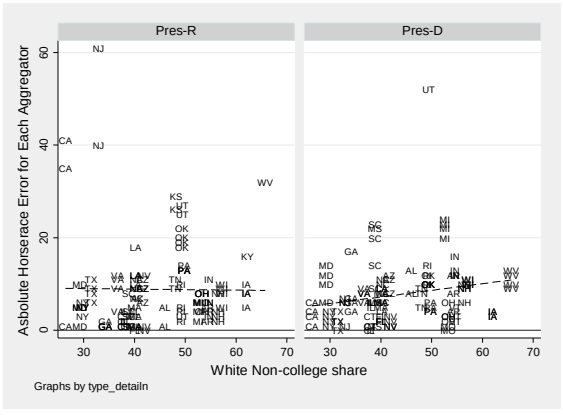
Figure A.12 Signed Error Averaged Across Aggregators and for Each Aggregator by Percent of White Noncollege Voters, by Contest

The left panel of Figure A.13 examines the absolute horserace error for Republican and Democratic primary contests separately, averaged across the aggregators. Here, as with the average signed horserace error, there is no correlation overall between the share of white non-college voters and the absolute horserace error averaged over all of the aggregators (r=0.11, p=0.40). Furthermore, there is no correlation between the average absolute error across the aggregators and the share of white non-college voters in either the Republican (r=0.03, p=0.86) or Democratic races (r=0.19, p=0.29).

When we instead examine the correlation between the share of white non-college voters and the absolute prediction errors for each of the aggregators, rather than the average across the aggregators, we see a similar pattern as for signed error – no correlation for the Republican contests (r=-0.01, p=0.91) and a significant positive correlation for the Democratic contests (r=0.21, p=0.02). That is, absolute errors were greater for Democratic contests with more white non-college voters, but not for Republican contests. This pattern is shown in the right panel of Figure A.13.



Absolute error averaged across aggregators



Absolute error for each aggregator

Figure A.13 Absolute Error Averaged Across Aggregators and for Each Aggregator by Percent of White Noncollege Voters, by Contest

We now consider all of these factors simultaneously in a hierarchical linear model with the 230 predictions nested within 36 states across both the Republican and Democratic contests, using states as random effects using Stata's `xtmixed` command. We then add the variables examined above (except for whether a correct prediction was made) to account for any confounding between the predictor variables.

Table A.16 presents regression coefficients and standard errors predicting both the signed horserace error and absolute horserace error. For signed errors, the intraclass correlation coefficient (ICC) is 0.374, indicating that 37.4% of the variance in the signed errors is within states (across predictions); for absolute errors, the ICC is 0.295, indicating that 29.5% of the variance in absolute errors is within the states (across aggregator predictions), and the remainder is between the states.

To be consistent with the multivariate discussion in the state-level polling section, we will focus our interpretation of the multivariate models on the absolute horserace errors rather than signed errors. Overall, Democratic contest predictions fared differently from Republican contest predictions, with Democratic contest predictions being more accurate overall (a negative coefficient in absolute error indicates lower errors, or more accurate). Although the association between absolute horserace prediction errors and the percent of white non-college voters in the state was fairly flat for the Republican contests, errors *increased* in Democratic contests for states with a higher share of white non-college educated voters. Said another way, the Democratic contest predictions were less accurate in states with higher concentrations of white non-college voters.

As seen in the bivariate analyses, absolute errors were greater in states where the winner took a larger percentage of the vote. Additionally, states where there were more predictions available had smaller absolute errors. The FiveThirtyEight Polling Average fared worse than the other aggregator predictions, but these only occurred where there were few polls available.

Table A.16 Regression Coefficients and Standard Errors (in parentheses) from Hierarchical Linear Model Predicting Signed Error and Absolute Error, Aggregators Only

	Signed horserace error			Absolute horserace error		
	Null	Main effects	Interaction	Null	Main effects	Interaction
Democratic contests		5.69****	5.52****		-3.70****	-3.49****
		(1.24)	(1.24)		(1.00)	(0.99)
% white non-college voter population		-0.10	-0.01		0.06	-0.05
		(0.14)	(0.15)		(0.10)	(0.11)
Democratic contests *% white non-college voter population			-0.16			0.18*
			(0.10)			(0.08)
% for winner		-0.48****	-0.47****		0.25****	0.24****
		(0.06)	(0.06)		(0.05)	(0.05)
Number of predictions		2.81*	3.04*		-4.19****	-4.54****
		(1.32)	(1.33)		(1.02)	(1.03)
Firm (reference=RCP)						
FiveThirtyEight Polling Average		-3.13	-2.65		8.65*	8.01*
		(4.56)	(4.53)		(3.65)	(3.60)
FiveThirtyEight Polls-Only		0.33	0.38		-0.02	-0.06
		(1.38)	(1.36)		(1.11)	(1.09)
FiveThirtyEight Polls-Plus		0.74	0.78		-0.25	-0.30
		(1.38)	(1.36)		(1.11)	(1.09)
Huffington Post		0.88	0.88		0.70	0.69
		(1.43)	(1.41)		(1.15)	(1.13)
Intercept	-4.95****	-7.68	-18.58****	8.74****	11.52*	13.22**
	(1.29)	(6.20)	(4.96)	(0.91)	(4.83)	(4.84)
Variance components						
SD State random effects	6.76****	7.15****	7.25****	4.57****	4.99****	5.22****
SD Residual variance	8.74	7.29	7.22	7.07	5.89	5.78
State ICC	0.374	0.491	0.502	0.295	0.418	0.449

Note: n=230 predictions in 36 states; *p<.05, **p<.01, ***p<.001, ****p<.0001; % White non-college population grand mean centered at 44.18879; Winner percent grand mean centered at 53.67931.

A.G Using Callback Studies to Look at the Stability of Vote Preference

With tracking data collected on independent samples of respondents, it is difficult to disentangle changes in sample composition from changes in individual voting intentions. Both Pew Research Center's American Trends Panel (ATP) and YouGov's Cooperative Congressional Election Study (CCES) recontacted those interviewed before the election after November 9. Table A.17 shows the relationship between pre-election voting intention and post-election vote report for registered voters who were interviewed in the October and November waves of Pew's ATP. Table A.18 performs the same analysis for YouGov's CCES. Data are not weighted and exclude respondents who did not take either wave, were undecided, or did not plan to vote.

The pattern of stability is similar in Tables A.17 and A.18. There was hardly any switching between the two major party candidates, though fewer Clinton supporters in the pre-election wave turned out to vote than Trump supporters. There were losses for the Libertarian and Green candidates: Johnson supporters broke toward Trump, while Stein supporters broke toward Clinton, but the overall numbers were relatively small. These data indicate that there was a small gain for Trump between the pre- and post-election interviews, but that the amount of switching was small compared to the swings seen in pre-election tracking surveys. This suggests that a substantial portion of the variation in pre-election tracking was compositional. Actual switches in voting intention were rare. Only about 1-in-20 respondents reported voting for a different candidate in the post-election interview from the one they supported in their pre-election interview. The net shift toward Trump was quite small and occurred mostly through differential turnout.

One cautionary note about this analysis, however, is that the possibility of time-in-panel effects. The ATP panelists were asked about their vote preference roughly monthly during 2016, and the CCES panelists were asked multiple times as well. There is some concern that answering the vote choice question again and again could have made the election more salient for these panelists than for typical adults. If the panelists were more engaged and more likely to have solidified their vote choice, then the data here would potentially under-state the real level of vote choice fluctuation among the electorate. It seems unlikely that time-in-sample effects would meaningfully undermine these data, but the possibility deserves mention.

Table A.17 Differences between Pre-election Vote Intention and Reported Vote (Pew ATP)

	Pre-election interview				
Post-election interview	Trump	Clinton	Johnson	Stein	Total
Trump	36.4	0.5	0.9	0.3	38.0
Clinton	0.2	46.0	0.5	0.5	47.2
Johnson	0.4	0.2	3.4	0.2	4.1
Stein	0.1	0.1	0.1	1.2	1.3
Did not vote	2.9	3.9	1.7	0.8	9.3
Total	40.0	50.6	6.5	2.9	100.0

Table A.18 Differences Between Pre-election Vote Intention and Reported Vote (YouGov CCES)

	Pre-election interview				
Post-election interview	Trump	Clinton	Johnson	Stein	Total
Trump	30.6	0.5	0.7	0.4	32.2
Clinton	0.4	35.1	0.5	0.5	36.5
Johnson	0.3	0.3	3.2	0.2	4.1
Stein	0.0	0.0	0.1	1.3	1.4
Did not vote	7.3	12.1	3.8	2.6	25.8
Total	38.5	48.2	8.2	5.1	100.0

A.H Adjusting on a More or Less Detailed Education Variable

Table A.19 compares the education distribution of registered voters in four samples (the October wave of the ATP, the combined ABC News/Washington Post daily tracking data, the CCES pre-election wave, and the combined Survey Monkey (SM) tracking data) with the 2012 CPS Voting and Registration Supplement. The data have been weighted using the post-stratification weight created by each survey organization. Except for the Pew ATP (which slightly over-represents registered voters without a high school degree), the samples have too few high school dropouts. The two online samples have collapsed "less than high school" and "high school graduates" into a single weighting cell, so they substantially underrepresent the number of persons in the lowest education category. This is offset by an over-representation of the next education category (high school graduates).

Table A.19 Weighted Percentage of Education Level in Five Samples of Registered Voters

	Education				
Survey	No HS Degree	HS Graduate	Some College	College Graduate	Post Graduate
CPS 2012	7.2	27.2	31.4	22.2	12.0
Pew Research Center	9.0	27.5	33.4	18.6	11.5
ABC News/Washington Post	6.1	25.1	23.0	32.1	13.7
CCES 2016	2.7	32.6	32.6	19.9	12.2
SurveyMonkey	2.6	30.1	31.7	19.0	16.7

In the 2016 election, there was a strong correlation between education and candidate preference for white votes. (According to the NEP Exit poll, Trump won whites without a college degree 66-29 compared to 48-45 among whites with a college degree. This is almost three times as large

a difference as in the 2012 Presidential election.) While this could potentially affect vote estimates, it does not appear that they were a significance source of error. A simple correction—post-stratifying the survey weights by 5 education categories—results in only small changes in the pre-election vote estimates (less than 0.4 percent) and no systematic improvement.

A.I Adjusting on Political Party Affiliation

In pre-election polls, there is the possibility of non-demographic sample skews, such as party identification. If a sample contains too many respondents of one party, it is almost certain to overestimate the vote for that party and weighting the sample by the population proportion of party identifiers could measurably improve the accuracy of vote estimates. Party identification (ID) weighting, however, is controversial. First, there is no widely accepted benchmark for party ID. Second, even if there was an authoritative source that could be used, party ID varies over time, so weighting to an out-of-date target could mask party ID swings and make the party-weighted estimates less, rather than more, accurate.

Even if the population distribution of party ID is not available and party ID changes over time, different distributions of party ID in data collected at the same period is evidence of partisan selection effects. Day-to-day variations in sample composition *could* be due to shifts in party ID, but daily shifts are implausibly large. Panel data on three category party ID exhibit a high degree of stability and showed no trend toward one party or the other during the 2016 campaign. For example, in Pew Research Center's American Trends Panel (ATP), 90 percent of Democrats and Republicans chose the same party ID in the post-election survey as in the pre-election survey. The instability was almost entirely to or from the independent category, with offsetting movements in either direction. Hardly anyone (27 out of 3,961 respondents) changed from one party to the other.

There is, however, evidence that differences between the partisan composition of different polling samples—even after demographic weighting—resulted in substantial differences in vote estimates. Figure A.14 shows the relationship between the (weighted) sample party ID distribution and daily vote estimates in two tracking polls (the online survey is SurveyMonkey and the phone tracking is ABC). Party ID is measured by the percentage of Democrats in the sample, less the percentage of Republicans. Vote intention is the percentage of voters intending to vote for Clinton, less the percentage intending to vote for Trump. The online survey had much larger sample sizes (over 5,000 most days, and over 10,000 per day for most of the final week of the campaign) than the phone tracking (about 264 respondents on the average day). As a consequence, the phone tracking is much noisier, but the pattern is very simple in both samples: an extra percent of Democratic party identifiers in the sample increases Democratic vote intention by about one percent. That said, it is important to note that neither SurveyMonkey nor ABC/Washington Post published estimates for single day interviewing. Their weighting protocols are not designed for that purpose. So in a sense, this analysis overstates how much of a problem daily variation is for such polling organizations.

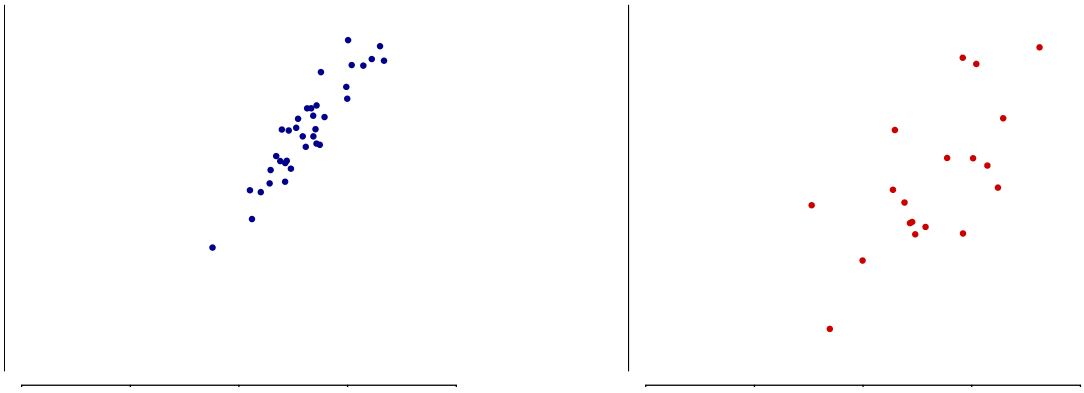


Figure A.14 Relationship Between Sample Composition of Party Identifiers and Vote Intention for an Online Survey (Left Panel) versus Live Phone Survey (Right Panel).
Source: Left panel is SurveyMonkey data; Right panel is ABC/Washington Post data
Note – Neither SurveyMonkey nor ABC/Washington Post published estimates for single day interviewing.

This is similar to the pattern found by Gelman and colleagues (2016) in 2012 Presidential election polling. On some days, the samples contained as much as 20 percent more Democrats than Republicans, while on other days there were more Republicans than Democrats. These partisan surges were associated with campaign events and gave the impression that there were large swings in voting intentions, despite the use of demographic weighting. In 2016, similar large swings were seen in both online and telephone tracking (as shown in the left panel of Figure A.15). On some days, Clinton led by as much as 15 percent, while on others Trump led by nearly as much. The size of these swings is much larger than could possibly be due to sampling variability. (The standard error of daily differences in lead is approximately 2 percent for the large online samples and about 8 percent for the smaller daily phone tracking samples.)

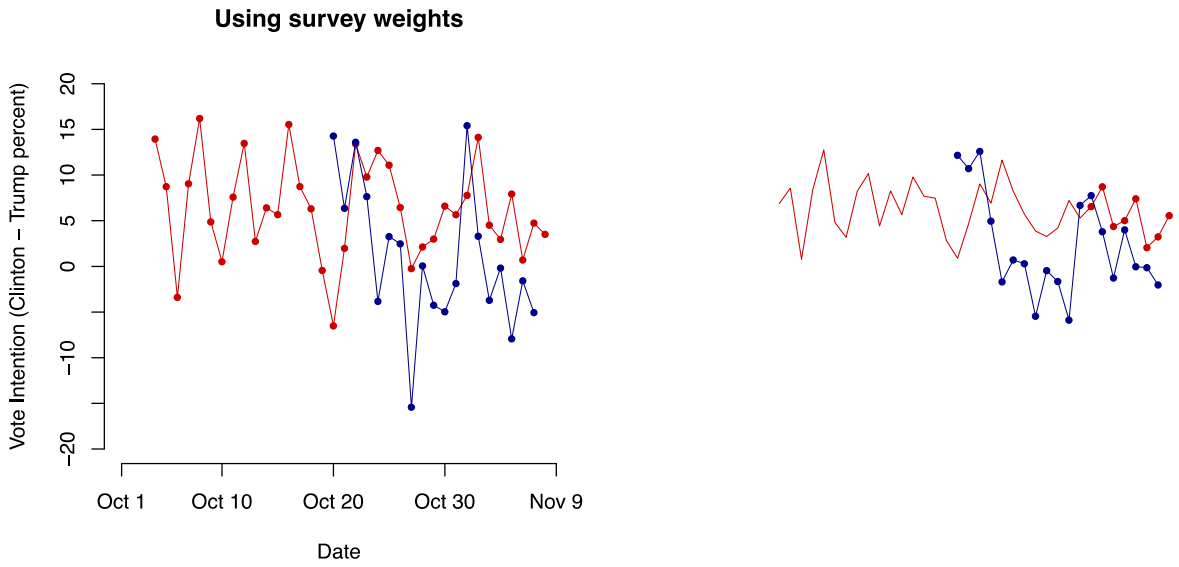


Figure A.15 Daily Clinton Lead Using Survey Weights (left) and Post-stratified by Party Identification (right). Notes – Online survey shown in red, phone tracking in blue. Neither SurveyMonkey nor ABC/Washington Post published estimates for single day interviewing.

The weights can be adjusted to reflect a constant distribution of party ID, rather than a dynamic party ID target. To explore this, we pooled each tracking sample to estimate the proportions of Democrats, Republicans, and Independents in the organization's tracking data. The base survey weights were then post-stratified by the proportions in each party ID group, so that the weighted proportions of Democrats and Republicans would be the same each day. Daily leads were computed using these weights and are shown in the right panel of Figure A.15. About half of the daily variation in the Clinton lead is removed by reweighting with a static party ID target. The online sample shows no evident trend over the last month of the campaign. The phone tracking data is noisier, but the daily differences are now in the range that could be explained by normal sampling variation.

While this analysis demonstrates that adjusting for party ID can reduce day-to-day variability in poll estimates, it is not clear that such an adjustment reduces bias. For at least one poll, adjusting on a close cousin – political ideology – did not help. According to Blumenthal (2016), SurveyMonkey's practice of weighting to match its own previous results for ideology had the effect of boosting Clinton's vote total in the final week of the campaign. Dropping the ideology smoothing would have reduced Clinton's margin in the final week from +6.0 to +4.8 (relative to a vote outcome of +2.1). It is also worth noting that the approach tested here (i.e., pooling all the tracking data collected by a firm during an election season) is not something that can be implemented in practice. In practice, the pollster only has data for the interviewing conducted to date, as opposed to the entire tracking series, which includes interviewing conducted in the future.

A.J 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 2016c). 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.

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. (2016), "The Latest Data and Methodological Information on How. SurveyMonkeY Measures Shifts in Voter Sentiment." Retrieved from https://blog.electiontracking.surveymonkey.com/2016/12/22/looking-back-at-2016-what-weve-learned-so-far/.

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.

Gelman, A., Goel, S., Rivers, D., and Rothschild, D. (2016), "The Mythical Swing Voter," *Quarterly Journal of Political Science*, 11(1), 103-130.

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/.

Silver, N. (2016b), "Pollsters Probably Didn't Talk to Enough White Voters Without College Degrees," FiveThirtyEight.com, December 1, 2016. Retrieved from https://fivethirtyeight.com/features/pollsters-probably-didnt-talk-to-enough-white-voters-without-college-degrees/.

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