2-2018

An Evaluation of the 2016 Election Polls in the United States

Courtney Kennedy
Pew Research Center, Washington, DC, ckennedy@pewresearch.org

Mark Blumenthal
SurveyMonkey

Scott Clement
Washington Post

Joshua D. Clinton
Vanderbilt University, josh.clinton@vanderbilt.edu

Claire Durand
University of Montreal, claire.durand@umontreal.ca

See next page for additional authors

Kennedy, Courtney; Blumenthal, Mark; Clement, Scott; Clinton, Joshua D.; Durand, Claire; Franklin, Charles; McGeeney, Kyley; Miringoff, Lee; Olson, Kristen M.; Rivers, Douglas; Saad, Lydia; Witt, G. Evans; and Wlezien, Christopher, "An Evaluation of the 2016 Election Polls in the United States" (2018). *Sociology Department, Faculty Publications*. 543.
http://digitalcommons.unl.edu/sociologyfacpub/543

Authors
Courtney Kennedy, Mark Blumenthal, Scott Clement, Joshua D. Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen M. Olson, Douglas Rivers, Lydia Saad, G. Evans Witt, and Christopher Wlezien

Published in *Public Opinion Quarterly,* Vol. 82, No. 1, Spring 2018, pp. 1–33.
doi:10.1093/poq/nfx047

Published February 3, 2018.

An Evaluation of the 2016 Election Polls in the United States

Courtney Kennedy,[1] Mark Blumenthal,[2] Scott Clement,[3]
Joshua D. Clinton,[4] Claire Durand,[5] Charles Franklin,[6]
Kyley McGeeney,[7] Lee Miringoff,[8] Kristen Olson,[9]
Douglas Rivers,[10] Lydia Saad,[11] G. Evans Witt,[12]
and Christopher Wlezien[13]

1 Courtney Kennedy is director of survey research at the Pew Research Center, Washington, DC, USA.
2 Mark Blumenthal is head of election polling at SurveyMonkey, Washington, DC, USA.
3 Scott Clement is polling director at the *Washington Post*, Washington, DC, USA.
4 Joshua D. Clinton is Abby and Jon Winkelried Chair of Political Science and codirector of the Center for the Study of Democratic Institutions at Vanderbilt University, Nashville, TN, USA.
5 Claire Durand is a professor of sociology at the University of Montreal, Montreal, Canada.
6 Charles Franklin is a professor of law and public policy and director of the Marquette Law School Poll at Marquette University, Madison, WI, USA.
7 Kyley McGeeney is senior director of survey methods at PSB, Washington, DC, USA.
8 Lee Miringoff is an associate professor of political science and director of the Marist Institute for Public Opinion at Marist College, Poughkeepsie, NY, USA.
9 Kristen Olson is an associate professor of sociology at the University of Nebraska–Lincoln, Lincoln, NE, USA.
10 Douglas Rivers is a professor of political science and senior fellow at the Hoover Institution at Stanford University, Stanford, CA, USA, and chief scientist at YouGov, Redwood City, CA, USA.
11 Lydia Saad is a senior editor for the Gallup Organization, Washington, DC, USA.
12 G. Evans Witt is CEO of Princeton Survey Research Associates International, Washington, DC, USA.
13 Christopher Wlezien is Hogg Professor of Government at the University of Texas at Austin, Austin, TX, USA.

Corresponding author — Courtney Kennedy, Pew Research Center, 1615 L St., NW, Suite 800, Washington, DC 20036, USA; *email* ckennedy@pewresearch.org

Abstract

The 2016 presidential election was a jarring event for polling in the United States. Preelection 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, there was a widespread perception that the polls failed. But did the polls fail? And if so, why? Those are among the central questions addressed by an American Association for Public Opinion Research (AAPOR) ad hoc committee. This paper presents the committee's analysis of the performance of preelection polls in 2016, how that performance compares to polling in prior elections, and the extent to which performance varied by poll design. In addition, the committee examined several theories as to why many polls, particularly in the Upper Midwest, underestimated support for Trump. The explanations for which the most evidence exists are a late swing in vote preference toward Trump and a pervasive failure to adjust for overrepresentation of college graduates (who favored Clinton). In addition, there is clear evidence that voter turnout changed from 2012 to 2016 in ways that favored Trump, though there is only mixed evidence that misspecified likely voter models were a major cause of the systematic polling error. Finally, there is little evidence that socially desirable (*Shy Trump*) responding was an important contributor to poll error.

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). Leading up to the election, three types of information widely discussed in the news media indicated that Democratic nominee Hillary Clinton was likely to win. First, polling data showed Clinton consistently leading the national popular vote, which is usually predictive of the winner (Erikson and Wlezien 2012), and leading, if narrowly, in Pennsylvania, Michigan, and Wisconsin—states that had voted Democratic for president six elections running. Second, early voting patterns in key states, particularly in Florida and North Carolina, were described in high-profile news stories as favorable for Clinton (Silver 2017a). Third, election forecasts from highly trained academics and data journalists declared that Clinton's probability of winning was about 90 percent, with estimates ranging from 71 to over 99 percent (Katz 2016).

The day after the election, there was a palpable mix of surprise and outrage directed toward 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 2016; Shepard 2016). The unexpected US outcome added to concerns about polling raised by errors in the

2014 referendum on Scottish independence, the 2015 UK general election, and the 2016 British referendum on European Union membership (Barnes 2016).

In the weeks after the 2016 US election, states certified their vote totals and researchers began assessing what happened with the polls. It became clear that a confluence of factors made the collective polling miss seem worse than it actually was, at least in some respects. The winner of the popular vote (Clinton) was different than the winner of the Electoral College (Trump). While such a divided result is not without precedent, the full arc of US history suggests it is highly unlikely. With respect to polling, preelection estimates pointed to an Electoral College contest that was less certain than interpretations in the news media suggested (Trende 2016; Silver 2017b). 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). Trende noted that his organization's battleground-state poll averages had Clinton leading by a very slim margin in the Electoral College (272 to 266), putting Trump one state away from winning the election. Relatedly, the elections in the three Upper Midwest states that broke unexpectedly for Trump (Pennsylvania, Michigan, and Wisconsin) were extremely close. More than 13.8 million people voted for president in those states, and Trump's combined margin of victory was 77,744 votes (0.56 percent). Even the most rigorously designed polls cannot reliably indicate the winner in contests with such razor-thin margins.

Even with these caveats about the election, a number of important questions surrounding polling remained. There was a systematic underestimation of support for Trump in state-level and, to a lesser extent, national polls. The causes of that pattern were not clear but potentially important for avoiding bias in future polls. Also, different types of polls (e.g., online versus live telephone) seemed to be producing somewhat different estimates. This raised questions about whether some types of polls were more accurate and why. More broadly, how did the performance of 2016 preelection polls compare to those of prior elections?

These questions became the central foci for an ad hoc committee commissioned by the American Association for Public Opinion Research (AAPOR) in the spring of 2016. The committee was tasked with summarizing the accuracy of 2016 preelection polling, reviewing variation by different poll methodologies, and assessing performance through a historical lens. After the election, the committee decided to also investigate why polls, particularly in the Upper Midwest, underestimated support for Trump.

The next section presents several of the main theories for why many polls underestimated Trump's support. This is followed by a discussion of the data and key metrics the committee used to perform its analyses. Subsequent sections of the paper present analyses motivated by the research questions posed here. The paper concludes with a discussion of the main findings and implications for the field.

Theories about Why Polls Underestimated Support for Trump

A number of theories were put forward as to why many polls missed in 2016.[1]

Nonresponse Bias and Deficient Weighting

Most preelection polls have single-digit response rates or feature an opt-in sample for which a response rate cannot be computed (Callegaro and DiSogra 2008; AAPOR 2016). While the link between low response rates and bias is not particularly strong (e.g., Merkle and Edelman 2002; Groves and Peytcheva 2008; Pew Research Center 2012, 2017a), such low rates do carry an increased risk of bias (e.g., Burden 2000). Of particular note, adults with weaker partisan strength (e.g., Keeter et al. 2006), lower educational levels (Battaglia, Frankel, and Link 2008; Chang and Krosnick 2009; Link et al. 2008; Pew Research Center 2012, 2017a), 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 less likely than other voters to accept survey requests. If survey response was correlated with presidential vote and some factor not accounted for in the weighting, then a deficient weighting protocol could be one explanation for the polling errors.

[1] The original committee report (AAPOR 2017) also discussed ballot-order effects. That discussion has been dropped in this paper because there was not strong evidence that such effects were a major contributor to polling errors in 2016. There remains an important debate about the possibility that ballot order affected the outcome of the presidential race in several states, including Michigan, Wisconsin, and Florida.

Late Deciding

The notion that preelection polls fielded closer to Election Day tend to be more predictive of the election outcome than equally rigorous polls conducted farther out has been well documented for some time (e.g., Crespi 1988; Traugott 2001; Erikson and Wlezien 2012). 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). Both Trump and Clinton had historically poor favorability ratings (Collins 2016; Yourish 2016). Unhappy with their options, some voters may have waited until the final week or so before deciding. Moreover, late deciders, being less anchored politically, tend to be more influenced by campaign events than voters deciding earlier (Fournier et al. 2004).

Misspecified Likely Voter Models

Constructing an accurate likely voter model is a tall order for even the most seasoned pollsters (Erikson, Panagopoulos, and Wlezien 2004). When turnout patterns diverge from recent elections, historical data can be unhelpful or even misleading. Voter turnout in 2016 differed from that in 2012 in ways that advantaged Trump and disadvantaged Clinton. Nationally, turnout among African Americans, the group most supportive of Clinton, dropped seven percentage points while turnout among Hispanics and non-Hispanic whites changed little, according to the Current Population Survey (CPS) Voting and Registration Supplement (File 2017). Furthermore, analysis by Fraga and colleagues (2017) indicates that the decline in African American turnout was sharpest in states such as Wisconsin and Michigan, which determined the outcome of the election. If pollsters designed their likely voter models around the assumption that 2016 turnout patterns would be similar to 2012, this could have led to underestimation of support for Republicans, including Trump. Such model misspecification could have been exacerbated by skews in the 2012 national exit poll (a popular source for turnout data) overstating turnout among young and nonwhite voters (McDonald 2007; Cohn 2016).

The "Shy Trump" Hypothesis (Reporting Error)

Controversy surrounding Trump's candidacy raised the possibility that some Trump voters may not have been willing to disclose their support for him in surveys. If a sizable share of Trump voters were reluctant to disclose their support for him, that could explain the systematic underestimation of Trump support in polls (e.g., Enns, Lagodny, and Schuldt 2017). Concern about the possibility of systematic misreporting of vote intention for or against a controversial candidate dates back decades. Studies examining this issue have tended to focus on elections in which either candidate race (Citrin, Green, and Sears 1990; Finkel, Guterbock, and Borg 1991; Traugott and Price 1992; Hopkins 2009) or gender (Hopkins 2009; Stout and Kline 2011) was a potential factor in polling error. In the 2016 presidential election, both race and gender were highly salient. Clinton was the first female major-party presidential nominee, and although both candidates were white, Trump's record on racially charged issues (e.g., housing discrimination, the Central Park Five, birtherism) and open support from white supremacists put race in the forefront of the campaign. However, a recent study suggests that the risk to polls from respondents intentionally misreporting vote choice has diminished considerably or disappeared entirely (Hopkins 2009).

Data and Methods

The committee used two types of data to evaluate the performance of polls and test the hypotheses listed above: respondent-level microdata sets and poll-level datasets. Given the large number of pollsters active during the election and the reality that all pollsters structure their microdata sets differently, the committee was selective in asking for microdata. ABC News/*Washington Post*, CNN, Marquette University, Michigan State University, Monmouth University, Pew Research Center, SurveyMonkey, USC/*LA Times*, and YouGov all provided microdata to the committee (Supplement Appendix A).

Poll-level datasets were compiled from FiveThirtyEight (via GitHub), HuffPollster, and RealClearPolitics. Those sources provide a few pieces of design information about each poll (e.g., pollster name, field dates, sample size, target population, and mode) in addition to the horserace estimates. For 2016 polls conducted close to Election Day, the committee supplemented those data with information about weighting, sample source, and the ratio of landline to cell phone interviews, where applicable. Adding these variables was done manually through searches of individual press releases, news stories, methodology reports, and pollster websites.

When design information about a poll was missing or unclear, the committee contacted individual pollsters to obtain the information. In all, the committee reached out to 59 different polling organizations, and 35 responded with at least partial information. Those who responded were generous with their time and information. Generally, noncooperation with the committee's requests did not have a noticeable impact on the work, with one exception. Nearly all pollsters using interactive voice response (IVR), sometimes called *robopolls*, did not respond to our requests. IVR polls represent a substantial share of the state-level polling conducted in 2016.

The committee selected two metrics to be the primary means of assessing poll performance. The absolute error on the poll margin was computed as the absolute value of the horserace margin (%Clinton minus %Trump) in the poll minus the same margin in the certified vote. For example, if a poll showed Clinton leading Trump by one point and she won by three 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 poll margin, which is computed in the exact same manner as the absolute error but without taking the absolute value. 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. The election polling literature offers several alternative metrics (e.g., Mosteller et al. 1949; Martin, Traugott, and Kennedy 2005), but the committee focused on the signed and absolute error on the margin because they are easily compared to past elections, they reflect how polls are actually discussed, and they are on a scale that a general audience can understand.

Analysis of the Performance of 2016 Polls Relative to Previous Elections

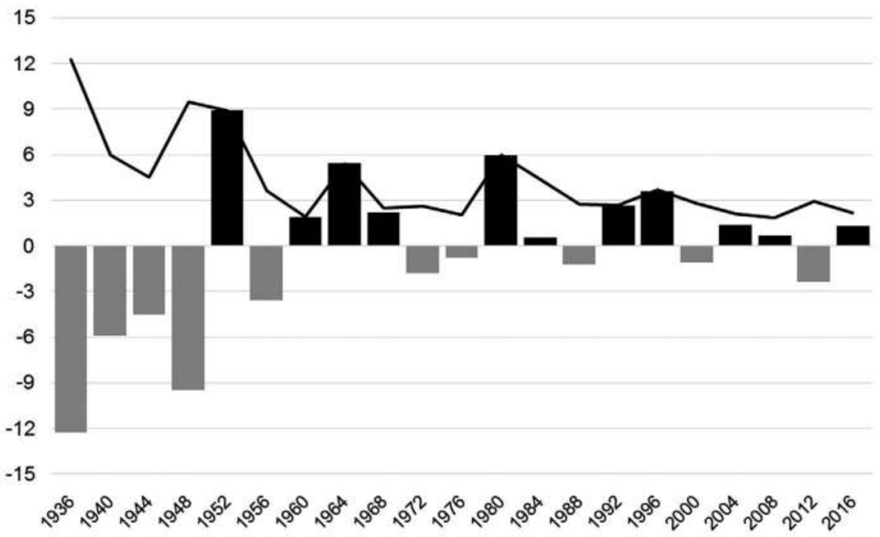
In the aftermath of the general election, many declared 2016 a historically bad year for polling. This first section examines the veracity of such conclusions by comparing the average error of 2016 preelection polls to that of polling in previous elections.

National General Election Polls[2]

As shown in Figure 1, national presidential polls in the 2016 general election were highly accurate by historical standards, resulting in small errors and correctly indicating that Clinton had a national popular vote lead close to her 2.1 percentage-point margin in the certified vote tallies. Final national polls in 2016 had an average absolute error of 2.2 percentage points, which is 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).

Examination of the average signed error in 2016 (1.3 percentage points) confirms that national polls tended to overestimate support for Clinton. Historically, it is not unusual for a frontrunner like Clinton to perform worse on Election Day than in the final polls (Erikson and Wlezien 2012). That said, the size and direction of the 2016 error contrasts with 2012, when polls underestimated Barack Obama's margin against Republican nominee Mitt Romney by 2.4 points. The average signed error in 2016 national polls was 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.

[2] This analysis includes polls that had a final field date within 13 days of Election Day (October 26 or later) and a starting date no earlier than October 16. 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 to 2012 and includes only final estimates.



	1936	1940	1944	1948	1952	1956	1960	1964	1968	1972	1976	1980	1984	1988	1992	1996	2000	2004	2008	2012	2016
Polls	1	1	1	1	1	1	1	2	2	3	3	4	6	5	6	9	10	16	19	25	22
ABS error	12.3	6.0	4.5	9.5	8.9	3.6	1.9	5.4	2.5	2.6	2.0	6.0	4.4	2.7	2.7	3.7	2.8	2.1	1.8	2.9	2.2
Signed error	-12.3	-6.0	-4.5	-9.5	8.9	-3.6	1.9	5.4	2.2	-1.8	-0.8	6.0	0.5	-1.2	2.7	3.6	-1.1	1.4	0.7	-2.4	1.3

Figure 1. Average error in vote margin in national presidential polls, 1936–2016. The line represents average absolute error. The bars represent average signed error (gray bars indicate overestimation of Republican vote margin; black bars indicate overestimation of Democratic vote margin). 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.

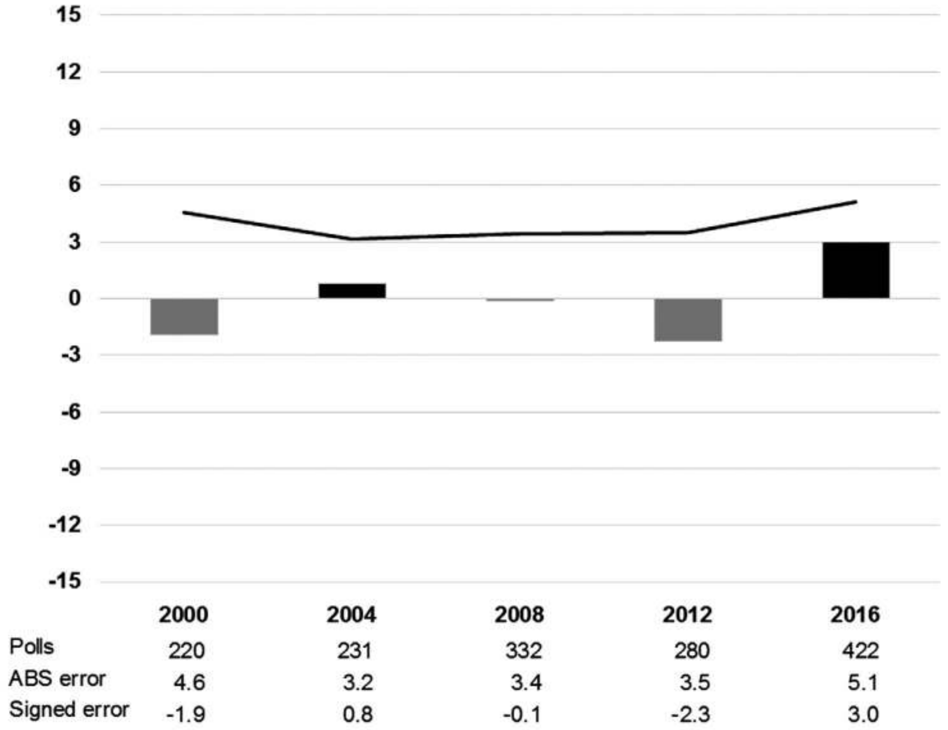
In recent elections, national polls have not consistently favored Republican or Democratic candidates. In 2016, national and state-level polls tended to underestimate support for Trump, the Republican nominee. In 2000 and 2012, however, general election polls clearly tended to underestimate support for 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.

State-Level General Election Polls

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 422 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 overestimating 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).

Both absolute errors and signed errors were smaller in battleground states than in non-battleground states (Supplement Appendix B). The average absolute error for the 206 battleground state polls was 3.6 points, compared with 6.5 points for the 216 polls in non-battleground states. The polls in non-battleground states underestimated Trump's vote margin against Clinton by 3.6 points on average (signed error); the underestimation of Trump's standing was 2.3 points in battleground states.

While the absolute errors tended to be lower in the more competitive states, underestimation 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



	2000	2004	2008	2012	2016
Polls	220	231	332	280	422
ABS error	4.6	3.2	3.4	3.5	5.1
Signed error	-1.9	0.8	-0.1	-2.3	3.0

Figure 2. Average error in vote margin in state presidential polls, 2000– 2016. The line represents average absolute error. The bars represent average signed error (gray bars indicate overestimation of Republican vote margin; black bars indicate overestimation of Democratic vote margin). Source.— Figures for 2000 to 2012 computed from data made public by FiveThirtyEight.com.

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

showing Clinton ahead by between two and 12 points in the final two weeks before she narrowly lost the state by less than one point. Similarly, before Clinton lost Pennsylvania and Michigan, polls in those states showed her with roughly three-point leads on average, which led to average signed errors of 4.2 and 3.5 points in those respective states. Underestimation of support for Trump was smaller in Florida, Arizona, and Georgia, while polls in Colorado and Nevada tended to overestimate his support.

The Performance of Primary and Caucus Polls

The 2016 presidential primary polls generally performed on par relative to past elections. Table 1 provides a summary. This analysis is based on all publicly released state-level 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. The polls correctly pointed to the winner in 86 percent of the 78 primaries with poll data available. The average absolute error across all primary polls reviewed was 9.3 points, not dramatically different, though slightly higher than errors in primary polls from recent elections. Supplement Appendix C provides additional analysis of the performance of primary and caucus polls.

Differences in Poll Accuracy by Survey Design

One limitation of aggregate analysis of polling errors is that it glosses over potentially important variation in performance by poll design. Given the diversity of designs currently in use, evidence of such variation would potentially be informative for election survey researchers and consumers. Many pollsters continue to use live telephone interviewing with random-digit-dial (RDD) samples of landlines and cell phones in the United States. Other pollsters conduct 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. 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).

The committee 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). These variables were coded for all national preelection surveys and battleground state surveys conducted in the final 13 days of the general election. The data are summarized in Figure 3. In terms of mode, national polls were twice as likely to be conducted by live telephone as battleground state polls (36 versus 18 percent, respectively). Battleground state polls were about twice as likely to be conducted using some form of IVR as national polls (41 versus 18 percent, 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. Sample 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 2.8 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. OLS regression analysis controlling for the potentially confounding effects from the specific contest polled and

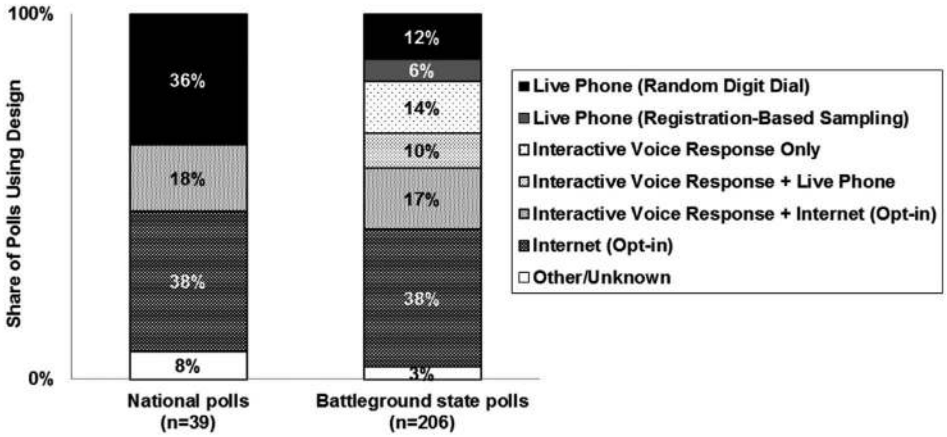


Figure 3. Design of 2016 general election polls conducted in final 13 days. The Franklin Pierce and Data Orbital polls, which were conducted by live telephone and had ambiguous statements about sample source that suggested RDD (but were not totally clear), are coded as live phone (RDD).

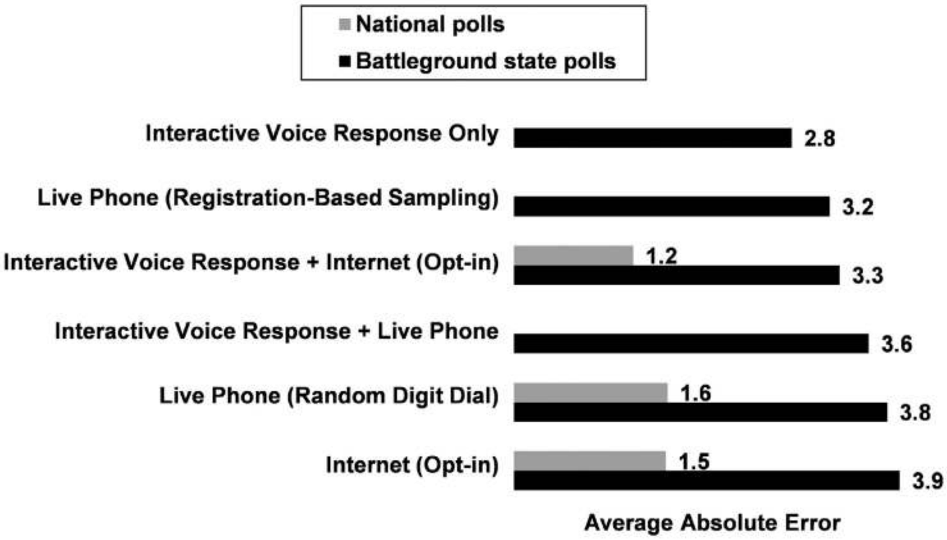


Figure 4. Average absolute error for 2016 general election polls, by design. Figures based on polls conducted during the final 13 days. Sample sizes for this analysis are small, and the effects from mode and sample source are to some extent confounded with house effects. National poll averages are based on seven polls (IVR+internet), 14 polls (live phone RDD) and 15 polls (internet opt-in). Battleground-state poll averages are based on 29 polls (IVR), 13 (live phone RBS), 34 polls (IVR+internet), 20 polls (IVR+live phone), 25 polls (live phone RDD), and 78 polls (internet opt-in).

number of days from Election Day corroborated the bivariate finding that polls using IVR tended to have less error in the 2016 general election (Supplement Appendix B).

The fact that IVR-only polls did relatively well is interesting in light of federal regulations dictating 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 may increase the risk of bias.[3]

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, undercoverage 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 (Online Supplement Appendix B) shows that IVR-only polls fared worse than other modes in both 2008 and 2012—elections in which Democratic turnout was relatively high. This suggests that the IVR results in 2016 may be an election-specific phenomenon related to the particular turnout patterns that year.

Evidence for Theories about Why Polls Underestimated Trump's Support

This section focuses on testing the major theories about why many general election polls underestimated support for Donald Trump.

[3] A review of methodological reports for IVR polls found that the commonly held assumption that such polls only dial landlines (Cassino 2016; Clinton and Rogers 2013; Cohn 2014; Enten 2012; Jackson 2016; Pew Research Center 2016) is not always correct. At least two pollsters described their methodology as IVR and yet reported that a noticeable share (10 to 25 percent) of their completed interviews were with cell phones.

Weighting: Education Was Strongly Correlated with Survey Participation and Presidential Vote

One hypothesis about 2016 polling errors is that pollsters did not interview enough white voters without a college degree (Silver 2016). Indeed, numerous studies have shown that adults with less formal education tend to be underrepresented in surveys (Battaglia, Frankel, and Link 2008; Link et al. 2008; Chang and Krosnick 2009; Pew Research Center 2012, 2017a). Generally speaking, this well-established education skew need not bias 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 underrepresentation of less educated adults in their weighting, 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.

Why would overrepresentation of college graduates 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 (right-hand panel in Figure 5). Historically, that has not been the case. In other modern US 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 5, in 2012 both the least-educated and most-educated voters broke heavily

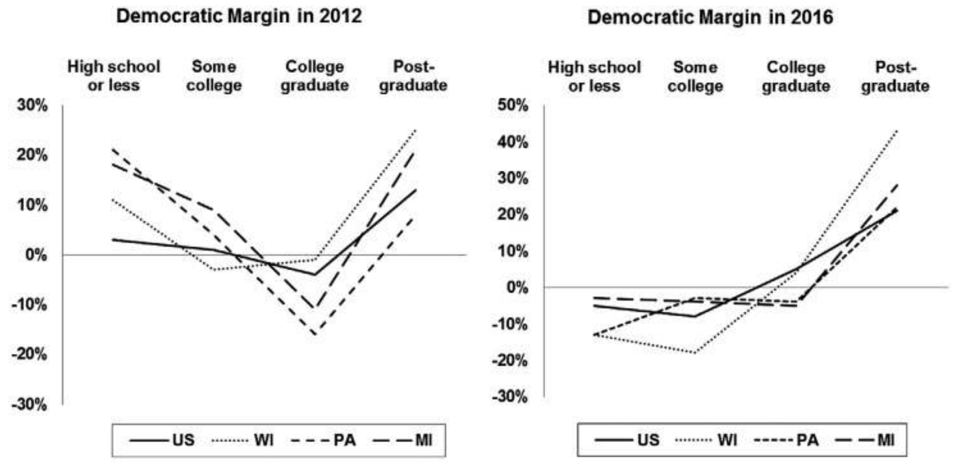


Figure 5. Democratic presidential vote margin in 2012 and 2016 by voter education level and geography. Source.—NEP national exit poll (2012, 2016).

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. Thus, the effects of weighting for education (or not) differ in 2012 versus 2016.

In 2012, the postgraduate voters who are likely to be overrepresented in polls that are not adjusted for education voted in much the same way as the low-education voters that such polls underrepresent. By contrast, in 2016, highly educated voters were poor proxies for the voters at the lowest education levels nationally and in the pivotal states in the Upper Midwest.

Following the election, two different state-level pollsters who had not adjusted for education in their preelection estimates reweighted their data to account for education. Both pollsters found that adjusting for education meaningfully improved their polls' accuracy by reducing estimates of Clinton support. The final University of New Hampshire (UNH) poll had Clinton leading by 11 points. She ultimately won by a razor-thin 0.4-point margin. According to UNH poll director Andrew 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 6, 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.

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.[4] She ultimately lost that contest by a 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

[4] 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 6.

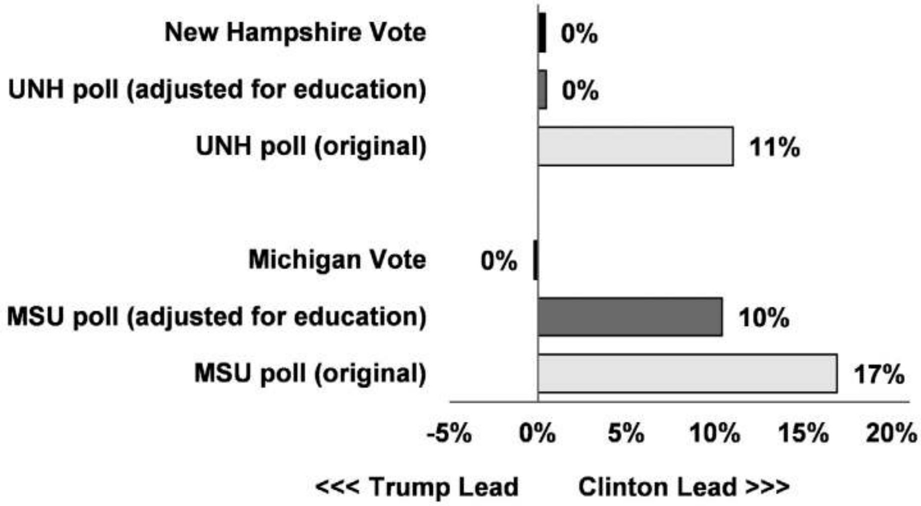


Figure 6. Poll estimates with and without weighting adjustment for education, relative to 2016 presidential vote outcome. Source.—University of New Hampshire poll conducted November 3–6, 2016, with 707 likely voters. Michigan State University poll conducted September 1–October 30, 2016, with 743 likely voters.

early, with most interviews completed before mid-October. Thus, the MSU poll largely missed what appears to be a significant, late shift in support to Trump (Blake 2016). As discussed below, 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.

To better understand the scope of this issue, the committee investigated how many polls in key races did or did not adjust for respondent education in their weighting. This effort required manual investigation and coding of each poll, so it was performed on a subset of the state contests (FL, OH, MI, NC, PA, WI) in addition to national polls. 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) were considered.[5]

[5] This yielded an analytic dataset with 102 polls. Despite outreach efforts to individual pollsters, we were unable to determine whether 17 of these polls had adjusted on education. Most of these undetermined 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. This imputation rule may be incorrect for a handful of polls, but based on the data that are present, it is highly unlikely that the imputation rule is wrong for a meaningful number of polls in this analysis.

Table 2. 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

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 2 shows that most state-level polls did not adjust for education in weighting, whereas about half of national polls did. In Michigan, under one-fifth of 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.

One contributor to this lack of accounting for education in weighting adjustments is that in 2016 the 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. Voter file samples provide pollsters with 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. Education, however, is not on the voter file and is generally absent from the weighting protocols of polls sampling off the voter file. 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 did not do so.

It appears that a number of IVR pollsters who sampled from voter files did not even measure respondent education. Table 3 shows that in Michigan, Pennsylvania, and Wisconsin only about half of the IVR pollsters were measuring respondent education, based on their press releases (which show gender, race, and other demographics).

We use the CPS Voting and Registration Supplement as benchmark data for the demographic profile of the voting electorate in 2016. The

Table 3. Share of college graduates in interactive voice response polls relative to the Current Population Survey in three states

Michigan		*Pennsylvania*		*Wisconsin*	
CPS benchmark	*38%*	*CPS benchmark*	*36%*	*CPS benchmark*	*37%*
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				

Benchmark data are weighted, filtered on self-reported voters, and come from the November 2016 Current Population Survey Voting and Registration Supplement. Election poll 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.

comparison shows that IVR polls overrepresented college graduates by at least 10 percentage points in critical Upper Midwest states (Table 3). Given that higher education levels were strongly associated with support for Clinton, the overrepresentation of more highly educated voters and not accounting for education in weighting contributed to errors in these states.

Did Polls Underrepresent Staunchly Pro-Trump Areas?

One question raised by the analysis above is whether the education imbalance was the only prevalent nonresponse bias in 2016 polls—or were there more? Nonresponse bias is notoriously hard to test, but one ecological analysis was possible. This analysis leveraged 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 underrepresented in polls. For example, if the Census shows that 13 percent of Americans live in staunchly pro-Trump areas, but polls estimate that only 9 percent of Americans live in those same areas, that would be evidence that polls were, indeed, systematically missing Trump supporters. However, there was no evidence to that effect. The results are presented in Table 4.

Since there was no obvious, definitive way to define a "staunchly pro- Trump" area, three definitions were tested. The definition used in the first row of the table identifies counties in which Trump won

Table 4. Estimates of the share of US adults living in staunchly pro-Trump counties

		Share of the US 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%

The Census figures are based on people of all ages. 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.

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, the 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 that Pew Research Center conducted in 2016 political polling.[6]

If the polls systematically missed 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

[6] As stated in the footnote of Table 4, 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. Analyses indicated that the discrepancy did not confound the comparison in a noticeable way. While it seemed possible that rural and other staunchly pro-Trump areas skew slightly older than other parts of the country, there was no empirical evidence of that. For example, the predominantly rural and overwhelmingly pro-Trump states of Oklahoma and Wyoming represented equal shares of the entire US population (1.2 and 0.2 percent, respectively) and the US adult population (also 1.2 and 0.2 percent, respectively). Consequently, this small discrepancy has no meaningful impact on the results or conclusions in this analysis.

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 overrepresented.

These findings do not rule out the possibility that differential nonresponse was a factor in 2016. It is possible that 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 underrepresentation of pro-Trump areas, does not show evidence of bias.

Likely Voter Modeling

The 2012 voting electorate was not a particularly good model for the 2016 voting electorate in key states. While the change in turnout may explain some of the polling error in 2016, just how much is difficult to quantify. Some pollsters lean heavily on the assumption that the past election is the best possible model of the coming election, but others do not (Supplement Appendix D discusses various approaches to likely voter modeling).

One 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 nonvoters 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. 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.

Very few surveys whose results were made public in 2016 sampled from voter lists in a way that readily facilitates validation. One

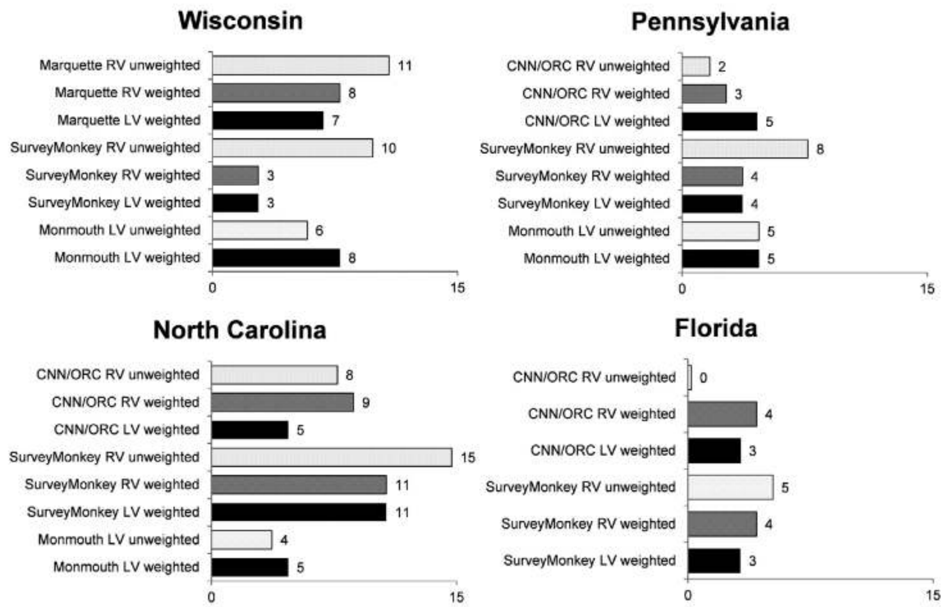


Figure 7. Signed error on 2016 presidential vote margin by poll, level of modeling, and state. The Marquette University poll was fielded October 27–30, 2016. The SurveyMonkey polls were all fielded November 1–7, 2016. The Monmouth University polls were fielded October 15–18, 2016 (Wisconsin), October 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).

example involves a series of polls conducted for the College of William and Mary by Target Smart, a Democratic-aligned data firm. A postelection analysis found that their respondents who actually voted were more likely to support Trump than respondents who did not vote. Excluding the validated nonvoters moved their estimate of Trump's percentage of the two-party (Trump and Clinton) vote in Ohio from 51 to 53 percent. Trump received 54 percent of the two-party vote in that state. Another way to assess the effect of likely voter modeling on accuracy is with 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 7 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.[7] The higher the value, the more the estimate overstated support for Clinton, relative to the election outcome. It is important to note that several of the polls included in the analysis were fielded more than two weeks out from Election Day and were not intended to be a final projection of the contest.

The results point to inconsistent effects from weighting and likely voter modeling across polls. In the SurveyMonkey polls, conducted online with op-tin sample, the weighting clearly helped improve accuracy. The likely voter model, however, tended to have little effect on SurveyMonkey's estimates for the states examined. 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 by several percentage points, making the final figures too Democratic. Non-Hispanic blacks constituted 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.

In Wisconsin, statistically adjusting the data slightly helped the Marquette University poll and slightly hurt the Monmouth University poll. In both cases, additional analysis revealed that the 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 improve accuracy while weighting up non-Hispanic blacks appears to have reduced accuracy. This result

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

comports with reporting of turnout patterns in the election. Overall, the analysis shows that postsurvey statistical adjustment reduced error in these polls, and more specifically reduced overestimation of Clinton support—by about two 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. Thus, statistical adjustment, in itself, does not appear to be an important cause of error in the polls broadly speaking.

Late Deciding: Evidence from National Exit Poll Data

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 in key battleground states and, to a lesser extent, at the national level. As reported by Blake (2016), the National Election Pool (NEP) exit poll conducted by Edison Research showed a substantial advantage for Trump among voters deciding their presidential choice 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 that 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. The pattern was not as strong nationally. The results are presented in Table 5.

These results suggest 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 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 happen days or weeks later on November 8), not a fundamental problem with their ability to measure public opinion at a given moment.

Table 5. 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%

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

Late Deciding: Evidence from a Callback Study

One limitation of the exit poll data is that they rely on people's recall about when they made a decision, which is prone to measurement error. A callback design, such as that employed in the Pew Research Center's callback telephone study, can also be informative about late decision-making, while avoiding reliance on recall. The Pew 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 from November 10 to 14, 2016. Only respondents who self-reported having voted were eligible to complete the postelection re-interview (n = 1,254).

In the callback study data, changes in vote preference manifest as discrepancies between pre- and postelection responses. It is also possible that *Shy Trump* responding would manifest the same way. Some 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 postelection interview. If poll respondents said they were undecided before the election and then said in November that they voted for Trump, the explanation could be either that they truly were undecided preelection or that they intentionally misreported as undecided. For some voters, the truth may fall somewhere in between.

The cross-tabulation of callback respondents' preelection and postelection responses is shown in Table 6. Cases on the downward, left-to-right diagonal represent respondents who answered the presidential

Table 6. Comparing individuals' pre- and postelection responses to presidential vote

	Reported vote				
Preelection vote preference	*Voted for Clinton*	*Voted for Trump*	*Voted for other candidate*	*Don't know 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%	
Don't know or 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 they voted in the general election. Estimates are unweighted.

vote question the same way before and after the election. About nine in 10 respondents (89 percent) answered consistently, while 11 percent reported doing 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, gave different pre- versus postelection responses (a result highly similar to Durand, Blais, and Vachon [2001]).

What is notable about the 2016 data is how the inconsistent responders voted. Figure 8 shows the presidential vote margin among respondents who gave inconsistent pre- versus postelection responses in callback studies since 2000. Typically, those who admit changing their minds break about evenly between the Republican candidate and the Democratic candidate. In 2016, by contrast, inconsistent responders in the Pew study voted for Trump by a 16-point margin. That is more than double the second largest margin in the time series (+7 points for George W. Bush in 2000).

The net effect on an election estimate based on such a preelection poll would be an error of just under two percentage points in underestimating support for Trump. Clinton's estimated national popular vote lead based on the registered voters in this study before the election was six percentage points, and her national lead based on those same individuals' post-election responses was four points. In addition, a small percentage of those screened for the postelection callback survey reported not voting (about 8 percent, n = 104). Clinton led Trump 44 to 27 percent among those who reported not voting. Thus, nonvoting hurt Clinton slightly more than it hurt Trump among this small sample.

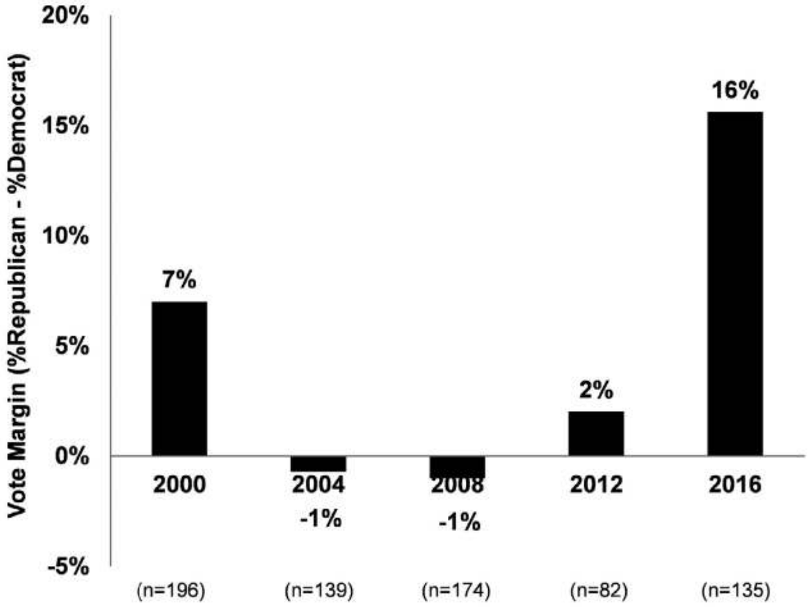


Figure 8. Vote margin (% voted for Republican candidate – % voted for Democratic candidate) among callback respondents giving inconsistent pre- vs. postelection responses. Data are from Pew Research Center RDD callback studies.

"Shy Trump" Reporting: Comparing Self- Versus Interviewer-Administered Modes

Is there evidence for the *Shy Trump* theory? If there was indeed a strong social desirability bias against expressing support for Trump, interviewer-administered polls (e.g., live phone) should record lower levels of Trump support than self-administered polls (e.g., IVR or online). For this test, all published polls conducted from September 1 to Election Day were examined. Figure 9 shows the national trend in Trump support, by mode, using a local regression estimation. It illustrates that estimates produced by live telephone polls were similar to those produced by self-administered online polls.

However, these aggregate effects may be due to other features of the polls than just mode of administration. To better isolate an effect from a mode, a regression analysis was conducted that controlled for length of field period, tracking poll versus nontracking poll, likely voter (LV) versus registered voter (RV) estimate, and change over time (Supplement Appendix E). The results show that self-administered online polls and interviewer-administered phone polls both recorded lower levels of support for Trump than IVR polls.

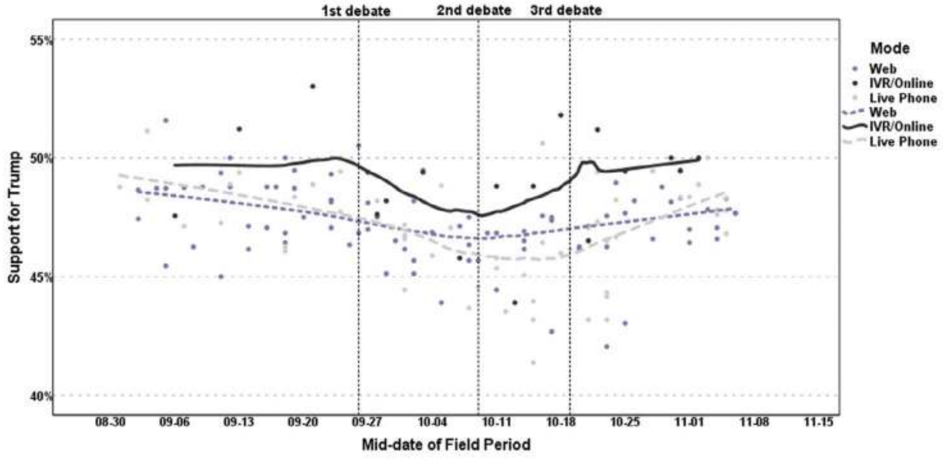


Figure 9. Support for Trump (on the sum of the two major party candidates) by mode. Each point represents a poll estimate positioned at the midpoint of the field period. Lines represent Loess estimates of change over time using Epanechnikov .65 estimation. IVR refers to interactive voice response. © C. Durand, 2016.

The finding that live telephone surveys did not consistently underestimate 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, this analysis offers no compelling evidence for the *Shy Trump* theory.

"Shy Trump" Reporting: Experiments Testing the Effect of Mode on Trump Support

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 (2017b) 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 eight percentage points more support for the conservative position online relative to on the phone. Taken together, these experiments suggest that there may have been some self-censoring of support for Trump in preelection polls, but the effect was likely to have been relatively small in size.

"Shy Trump" Reporting: Trump's Performance Relative to Republican Senate Candidates

A different way to test the *Shy Trump* hypothesis 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 at the individual-poll level, Trump should outperform his poll number by a larger margin than the Republican Senate candidate did.

Testing this theory involved using 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. Of the 34 Senate contests in 2016, eight were held in states where the presidential vote margin was less than five percentage points (AZ, CO, FL, NH, NV, NC, PA, WI). The analytic dataset of 66 polls included the final polls for these eight states and, for each state, only the last poll conducted by each firm. Overperformance is defined as the signed difference between the final vote margin and the poll margin, where the margin is the Republican vote minus the Democratic vote.[8,9] The central question is whether Trump tended to outperform 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, no support exists for that hypothesis. In the 24 live telephone polls analyzed, Trump beat his poll estimate by 1.4 percentage points on average, and the Republican Senate candidate beat his or her poll estimate by a nearly identical 1.3 percentage points on average. A similar, independent analysis reached the same general conclusion (Enten 2016).

In fact, not only did Trump outperform poll estimates, so did most Republican candidates in competitive Senate and House races. This pattern is evidenced by the fact that all of the values in the first two columns are positive. This finding is suggestive of systematic underestimation not just of support for Trump but of Republican candidates more generally. Indeed, Republican candidates for the US 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 underestimate support for Republican candidates writ large in 2016 suggests that polling errors were not caused by socially desirable reporting.

[8] For example, consider a Wisconsin poll showing the Senate race margin at –1 (44 percent for Johnson, the Republican, and 45 percent for Feingold, the Democrat) and the presidential margin at –6 (38 percent for Trump and 44 percent for Clinton). The actual elections in Wisconsin went +3.4 for Johnson and +0.7 for Trump. In this analysis, Johnson overperformed that poll by 3.4 – (–1) = +4.4 points, and Trump overperformed by 0.7 – (–6) = +6.7 points. Comparatively speaking, Trump overperformed the poll by 6.7 – 4.4 = 2.3 points more than the Republican Senate candidate did.

[9] "Overperformance" also can be defined 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. However, vote share was not a suitable framework. Due to the fact that polls feature undecided voters and tended to overestimate support for third-party candidates, both Donald Trump *and* Hillary Clinton generally "overperformed" relative to their estimated vote share in polls.

Table 7. Trump's overperformance of polls relative to Republican Senate candidates in battleground states

	Average overperformance (Vote margin – Poll margin)			
Type of poll	*Senate Rep. candidate*	*President Rep. candidate*	*Avg. difference (Pres. error – Sen. error)*	*Polls*
Live phone	1.3%	1.4%	0.0%	24
Online	4.5%	3.2%	–1.3%	17
Interactive voice response (alone or with online sample)	2.7%	1.8%	–0.9%	22
Other	7.7%	3.9%	–3.8%	3

"Shy Trump" Reporting: Effects of Interviewer Characteristics on 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, then 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. Because survey respondents are not randomly assigned to interviewers, statistical models are required to estimate the effects of interviewer race and sex on respondent vote preferences. In multivariate modeling controlling for basic respondent demographics (gender, race/ethnicity, education), there was no residual effect from interviewer race or gender on vote preference (Supplement Appendix F). The lack of evidence for an effect from interviewer characteristics on how respondents answered the presidential vote question is inconsistent with expectations of the *Shy Trump* theory, though it is not dispositive in ruling it out.

Conclusions

The committee, commissioned by AAPOR, conducted an extensive investigation of the performance of preelection polls in 2016. While the general public reaction was that the polls failed, we found the reality to be more complex. Some polls, indeed, had large problematic errors, but many polls did not. In particular, the national polls were generally correct (with respect to the popular vote) and accurate by historical standards. The accuracy of primary polls was typical of recent elections. The most glaring problems occurred in state-level general election polling, particularly in the Upper Midwest.

The committee evaluated a number of different theories as to why so many polls underestimated support for Donald Trump. The explanations for which the most evidence exists are a late swing in vote preference toward Trump and a pervasive failure to adjust for overrepresentation of college graduates (who favored Clinton). In addition, there is clear evidence that voter turnout changed from 2012 to 2016 in ways that favored Trump and other Republicans, though there is only mixed evidence that misspecified likely voter models were a major cause of the systematic polling error. Despite widespread speculation, there is little evidence that socially desirable (*Shy Trump*) responding was a major contributor to poll error. Experimental studies suggest that such mismeasurement error may have contributed errors on the order of a percentage point or two, but several other tests of this theory found no effect. If there was a *Shy Trump* effect on responses, it does not appear to have been particularly large.

One encouraging result from the historical analysis is that there is no systematic bias toward one major party or the other in US polling. In 2016, national and state-level polls tended to underestimate support for Trump, the Republican nominee. In 2012 and 2000, however, general election polls clearly tended to underestimate 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.

One broader question raised by this investigation is whether the polling problems in 2016 could reoccur. At a high level, the 2016 election featured a number of unusual circumstances that are perhaps unlikely to repeat (e.g., both major party candidates being historically unpopular, split outcomes in the popular vote and Electoral College, nearly 14 million votes across three states breaking for a candidate by about one-half of one percentage point), but several structural weaknesses of polls are likely to persist. Errors in state polls like those observed in 2016 are not uncommon, even though 2016 was a particularly bad election for state polls. With shrinking budgets at news outlets to finance polling, there is no reason to believe that the quality of state polling is going to noticeably improve in the near future. Finally, a late swing in favor of one candidate (as appears to have occurred in 2016) is not something that pollsters can necessarily guard against, other than by fielding closer to Election Day.

Acknowledgments — The committee was supported by a number of excellent research assistants: Junjie Chen, Andrew Engelhardt, Arnold Lau, Marc Trussler, and Luìs Patricio Peña Ibarra. The authors are grateful for comments, feedback, and editorial help from AAPOR Council members, as well as *POQ* editors and reviewers. They also thank Joan Murphy for extensive copy editing on an earlier version of this report.

Supplementary data appendices area attached to this document archive record.

References

American Association for Public Opinion Research (AAPOR). 2009. “An Evaluation of the Methodology of the 2008 Pre-Election Primary Polls.” http://www.aapor.org/Education-Resources/Reports/Methodology-2008-Primary-Polls.aspx

American Association for Public Opinion Research (AAPOR). 2016. *Standard Definitions Final Dispositions of Case Codes and Outcome Rates for Surveys*, 9th ed.

American Association for Public Opinion Research (AAPOR). 2017. A*n Evaluation of 2016 Election Polls in the U.S.* http://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx

Barnes, Peter. 2016. *“Reality Check: Should We Give Up on Election Polling?”* BBC News, November 11. http://www.bbc.com/news/election-us-2016-37949527

Battaglia, Michael P., Martin R. Frankel, and Michael W. Link. 2008. “Improving Standard Poststratification Techniques for Random-Digit-Dialing Telephone Surveys,” *Survey Research Methods* 2:11–19.

Blake, Aaron. 2016. *"How America Decided, at the Last Moment, to Elect Donald Trump."* Washington Post, November 17.

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

Burden, Barry C. 2000. "Voter Turnout and the National Election Studies." *Political Analysis* 8:389–98.

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

Callegaro, Mario, and Charles DiSogra. 2008. "Computing Response Metrics for Online Panels." *Public Opinion Quarterly* 72:1008–32.

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

Chang, Linchiat, and Jon A. Krosnick. 2009. "National Surveys Via RDD Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73:641–78.

Cillizza, Chris. 2016. *"Winners and Losers from the 2016 Election."* Washington Post, November 9. https://www.washingtonpost.com/news/the-fix/wp/2016/11/09/winners-and-losers-from-the-2016-election/?tid=a_inl&utm_term=.cbb00e865a74

Citrin, Jack, Donald P. Green, and David O. Sears. 1990. "White Reactions to Black Candidates: When Does Race Matter?" *Public Opinion Quarterly* 54:74–96.

Clinton, Joshua D., and Steven Rogers. 2013. "Robo-Polls: Taking Cues from Traditional Sources?" *PS: Political Science & Politics* 46:333–37.

Cohn, Nate. 2014. *"Two Polls That Highlight the Challenges of Polling."* New York Times, October 14. https://www.nytimes.com/2014/10/15/upshot/two-polls-that-highlight-the-challenges-of-polling.html

Cohn, Nate. 2016. *"There Are More White Voters Than People Think. That's Good News for Trump."* New York Times, June 9. https://www.nytimes.com/2016/06/10/upshot/there-are-more-white-voters-than-people-think-thats-good-news-for-trump.html?_r=0

Collins, Eliza. 2016. *"Poll: Clinton, Trump Most Unfavorable Candidates Ever."* USA Today, August 31. http://www.usatoday.com/story/news/politics/onpolitics/2016/08/31/poll-clinton-trump-most-unfavorable-candidates-ever/89644296/

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

Dropp, Kyle. 2016. "How We Conducted Our 'Shy Trumper' Study." https://morningconsult.com/2016/11/03/shy-trump-social-desirability-undercover-voter-study/

Durand, Claire, Andre Blais, and Sebastien Vachon. 2001. "A Late Campaign Swing or a Failure of the Polls? The Case of the 1998 Quebec Election." *Public Opinion Quarterly* 65:108–23.

Easley, Jonathan. 2016. *"Pollsters Suffer Huge Embarrassment."* The Hill, November 9. http://thehill.com/blogs/ballot-box/presidential-races/305133-pollsters-suffer-huge-embarrassment

Enns, Peter K., Julius Lagodny, and Jonathon P. Schuldt. 2017. "Understanding the 2016 U.S. Presidential Polls: The Importance of Hidden Trump Supporters." *Statistics, Politics and Policy* 8:41–63.

Enten, Harry. 2012. *"The Other 2012 Election Contest: Which Pollster and Polling Method Will Win?"* The Guardian, October 31. https://www.theguardian.com/commentisfree/2012/oct/31/other-2012-election-contest-pollster-polling

Enten, Harry. 2016. *"Shy' Voters Probably Aren't Why the Polls Missed Trump."* FiveThiryEight. com, November 16. https://fivethirtyeight.com/features/shy-voters-probably-arent-why-the-polls-missed-trump/

Erikson, Robert S., Costas Panagopoulos, and Christopher Wlezien. 2004. "Likely (and Unlikely) Voters and the Assessment of Campaign Dynamics." *Public Opinion Quarterly* 68:588–601.

Erikson, Robert S., and Christopher Wlezien. 2012. *The Timeline of Presidential Elections: How Campaigns Do (and Do Not) Matter.* Chicago: University of Chicago Press.

File, Thom. 2017. *"Voting in America: A Look at the 2016 Presidential Election."* U.S. Census Bureau, May 10. https://www.census.gov/newsroom/blogs/random-samplings/2017/05/voting_in_america.html

Finkel, Steven E., Thomas M. Guterbock, and Marian J. Borg. 1991. "Race-of-Interviewer Effects in a Preelection Poll: Virginia 1989." *Public Opinion Quarterly* 55:313–30.

Fournier, Patrick, Richard Nadeau, Andre Blais, Elisabeth Gidengil, and Neil Nevitte. 2004. "Time-of-Voting Decision and Susceptibility to Campaign Effects." *Electoral Studies* 23:661–81.

Fraga, Bernard L., Sean McElwee, Jesse Rhodes, and Brian Schaffner. 2017. *"Why Did Trump Win? More Whites — and Fewer Blacks — Actually Voted."* Washington Post, May 8. https://www.washingtonpost.com/news/monkey-cage/wp/2017/05/08/why-did-trump-win-more-whites-and-fewer-blacks-than-normal-actually-voted/?utm_term=.d0e1e494748f

Groves, Robert M., and Emelia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72:167–89.

Hopkins, Daniel J. 2009. "No More Wilder Effect, Never a Whitman Effect: When and Why Polls Mislead about Black and Female Candidates." *Journal of Politics* 71:69–81.

Jackson, Natalie. 2016. *"Here's Why HuffPost Is Dropping Polls That Rely Only on Landlines."* Huffington Post, August 1. http://www.huffingtonpost.com/entry/landline-only-polls-huffpost-pollster_us_579f9b2ae4b08a8e8b5ee65e

Jacobs, Jennifer, and Billy House. 2016. *"Trump Says He Expected to Lose Election Because of Poll Results."* Bloomberg.com, December 13. https://www.bloomberg.com/politics/articles/2016-12-14/trump-says-he-expected-to-lose-election-because-of-poll-results

Katz, Josh. 2016. *"Who Will Be President?"* New York Times, November 8. https://www.nytimes.com/interactive/2016/upshot/presidential-polls-forecast.html

Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best, and Peyton Craighill. 2006. "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey." *Public Opinion Quarterly* 70:759–79.

Link, Michael W., Michael P. Battaglia, Martin R. Frankel, Larry Osborn, and Ali H. Mokdad. 2008. "A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) for General Population Surveys." *Public Opinion Quarterly* 72:6–27.

Martin, Elizabeth A., Michael W. Traugott, and Courtney Kennedy. 2005. "A Review and Proposal for a New Measure of Poll Accuracy." *Public Opinion Quarterly* 69:342–69.

McDonald, Michael P. 2007. "The True Electorate: A Cross-Validation of Voter Registration Files and Election Survey Demographics." *Public Opinion Quarterly* 71:588–602.

Merkle, Daniel, and Murray Edelman. 2002. "Nonresponse in Exit Polls: A Comprehensive Analysis." In *Survey Nonresponse*, edited by Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little, pp. 243–57. New York: Wiley.

Mosteller, Frederick, Herbert Hyman, Phillip J. McCarthy, Eli S. Marks, and David B. Truman. 1949. *The Preelection Polls of 1948: Report to the Committee on Analysis of Preelection Polls and Forecasts.* New York: Social Science Research Council.

Noelle-Neumann, Elisabeth. 1974. "The Spiral of Silence: A Theory of Public Opinion." *Journal of Communication* 24:43–51.

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

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

Pew Research Center. 2017a. "What Low Response Rates Mean for Telephone Surveys." May 15. http://www.pewresearch.org/2017/05/15/what-low-response-rates-mean-for-telephone-surveys/

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

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

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

Silver, Nate. 2017a. *"Why Early Voting Was Overhyped."* FiveThirtEight.com, January 26. https://fivethirtyeight.com/features/early-voting-was-a-misleading-indicator/

Silver, Nate. 2017b. *"The Real Story of 2016: What Reporters—and Lots of Data Geeks, Too— Missed about the Election, and What They're Still Getting Wrong."* FiveThirtEight.com, January 19. http://fivethirtyeight.com/features/the-real-story-of-2016/

Smith, Andrew. 2016. "UNH 2016 Election Polls." Paper presented at the New England AAPOR Chapter Election Postmortem.

Stout, Christopher T., and Reuben Kline. 2011. "I'm Not Voting for Her: Polling Discrepancies and Female Candidates." *Political Behavior* 33:479–503.

Traugott, Michael W. 2001. "Trends: Assessing Poll Performance in the 2000 Campaign." *Public Opinion Quarterly* 65:389–419.

Traugott, Michael W., and Vincent Price. 1992. "The Polls—A Review: Exit Polls in the 1989 Virginia Gubernatorial Race: Where Did They Go Wrong?" *Public Opinion Quarterly* 56:245–53.

Trende, Sean. 2016. *"It Wasn't the Polls That Missed, It Was the Pundits."* RealClearPolitics. com. https://www.realclearpolitics.com/articles/2016/11/12/it_wasnt_the_polls_that_missed_it_was_the_pundits_132333.html

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

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