AAPOR

AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH

Task Force on 2020 Pre-Election Polling:

An Evaluation of the 2020 General Election Polls

Josh Clinton (Chair), Vanderbilt University[1]

Jennifer Agiesta, CNN

Megan Brenan, Gallup

Camille Burge, Villanova University

Marjorie Connelly, AP-NORC Center

Ariel Edwards-Levy, CNN

Bernard Fraga, Emory University

Emily Guskin, Washington Post

D. Sunshine Hillygus, Duke University

Chris Jackson, Ipsos

Jeff Jones, Gallup

Scott Keeter, Pew Research Center

Kabir Khanna, CBS News

John Lapinski, University of Pennsylvania

Lydia Saad, Gallup

Daron Shaw, University of Texas

Andrew Smith, University of New Hampshire

David Wilson, University of Delaware

Christopher Wlezien, University of Texas

[1] The Task Force thanks Sarah Lentz (University of Pennsylvania), Rezwana Uddin (NBC News), Mellissa Meisels (Vanderbilt University), Sara Kirshbaum (Vanderbilt University), Ami Ikuenobe (University of Pennsylvania), and Grayson Peters (University of Pennsylvania) for outstanding research assistance. The data collection and analysis would not have been possible without their careful and exhaustive work on behalf of the Task Force. In addition to the Task Force members, we would also thank Eunji Kim (Vanderbilt University), Steve Rogers (St. Louis University), John Sides (Vanderbilt University), Alan Wiseman (Vanderbilt University), and seminar participants at the Center for the Study of Democratic Institutions at Vanderbilt University for thoughtful and constructive feedback on an earlier version.

EXECUTIVE SUMMARY

The November 3, 2020, presidential election was historic by many standards, most notably because it was conducted during a global pandemic that resulted in a record high proportion of voters casting their ballots early and by mail. The election also featured the highest rate of voter turnout in decades.[2]

Most national polls accurately estimated that President Joe Biden would get more votes than President Donald Trump nationally, but Biden's certified margin of victory fell short of the average margin in the polls at both the national and state levels. Polling overstated support for Biden relative to Trump, and Biden's 306-232 victory in the Electoral College was narrower than predicted by many election forecasters.[3]

The shadow of 2016 hung over the 2020 election. Presidential election polls were widely criticized for "getting it wrong" in 2016. Many believed the polls predicted the wrong winner although the 2016 national polls were remarkably accurate in estimating former Secretary of State Hillary Clinton's national popular vote margin. Still, polls in a handful of key states overstated Clinton's lead or underestimated Trump's lead, leading to widespread discounting of Trump's Electoral College chances by election observers and rendering his victory a shock to many. The state-level polling error became the focus of a task force the American Association for Public Opinion Research (AAPOR) convened in spring 2016 to evaluate the accuracy of that year's polls, which resulted in very specific methodological recommendations for political polling (Kennedy et al. 2016).[4]

In October 2019, the Executive Council of AAPOR proactively convened a task force to examine the performance of pre-election polls in the 2020 elections. The Executive Council appointed 19 members to the Task Force on 2020 Pre-Election Polling from industry, nonprofit organizations, media, and academia to ensure a diversity of opinions, approaches, and expertise.

The Task Force collected all publicly available poll results at the national and state levels for the purpose of evaluating 2020 polling error and the extent that polls overstated the Democratic-Republican margin in the 2020 general election. The main findings of that review are as follows.

[2] https://www.pewresearch.org/fact-tank/2021/01/28/turnout-soared-in-2020-as-nearly-two-thirds-of-eligible-u-s-voters-cast-ballots-for-president/

[3] Initial returns suggested that polls did even worse than they ended up doing after all of the votes were counted. The outcome was not decided on election night: During the ongoing pandemic, the process of counting early votes and mail-in votes after Election Day resulted in the initial election returns being more favorable for Trump than the final certified vote.

[4] https://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx

The 2020 Pre-Election Polling Error

- The 2020 polls featured polling error of an unusual magnitude: It was the highest in 40 years for the national popular vote and the highest in at least 20 years for state-level estimates of the vote in presidential, senatorial, and gubernatorial contests.[5] Among polls conducted in the final two weeks, the average error on the margin in either direction was 4.5 points for national popular vote polls and 5.1 points for state-level presidential polls.

- The polling error was much more likely to favor Biden over Trump. Among polls conducted in the last two weeks before the election, the average signed error on the vote margin was too favorable for Biden by 3.9 percentage points in the national polls and by 4.3 percentage points in statewide presidential polls.

- The polling error for the presidential election was stable throughout the campaign. The average error matched closely for polls conducted in the last two weeks, in the final week, and even in the final three days. The challenges polls faced in 2020 did not diminish as Election Day approached.

- Beyond the margin, the average topline support for Trump in the polls understated Trump's share in the certified vote by 3.3 percentage points and overstated Biden's share in the certified vote by 1.0 percentage point.[6] When undecided voters are excluded from the base, the two-candidate support in the polls understated Trump's certified vote share by 1.4 percentage points and overstated Biden's vote share by 3.1 percentage points.

- The overstatement of the Democratic-Republican margin in polls was larger on average in senatorial and gubernatorial races compared to the presidential contest. For senatorial and gubernatorial races combined, polls on average were 6.0 percentage points too favorable for Democratic candidates relative to the certified vote margin. Within the same state, polling error was often larger in senatorial contests than the presidential contest.

- Whether the candidates were running for president, senator, or governor, poll margins overall suggested that Democratic candidates would do better and Republican candidates would do worse relative to the final certified vote.

- No mode of interviewing was unambiguously more accurate. Every mode of interviewing and every mode of sampling overstated the Democratic-Republican margin relative to the final certified vote margin. There were only minor differences in the polling error depending on how surveys sampled or interviewed respondents. Regardless of whether respondents were sampled using random-digit dialing, voter registration lists, or online recruiting, polling margins on average were too favorable to Democratic candidates.

- On average, polls overstated the Democratic-Republican margin in states more supportive of Trump in 2016. In states Trump won by more than five points in 2016, the average signed error on the margin was 5.3 percentage points too favorable for Biden; on the other hand, in states Clinton won by more than five percentage points in 2016, the average signed error on the margin was 3.5 percentage points too favorable for Biden. Even after controlling for state-level differences in demographics and voting administration, the average signed error was larger in states that favored Trump in 2016.

[5] It was the largest in 40 years for the national popular vote, but the performance of the state-level presidential polls has only been tracked since 2000.

[6] This calculation includes undecided voters in the base.

Factors That Do Not Explain the Polling Error

Several proposed explanations can be ruled out as primary sources of polling error in 2020. Our analyses suggest the following.

- **Polling error was not caused by late-deciding voters voting for Republican candidates**. More voters voted prior to Election Day in 2020 than ever before and the number of undecided voters was relatively small. Only 4% of poll respondents, on average, gave a response other than "Biden" or "Trump" when asked by state-level presidential polls conducted in the final two weeks. Unlike in 2016, respondents deciding in the last week were as likely to support Biden as Trump, according to the National Election Pool exit polls.

- **Polling error was not caused by a failure to weight by education.** A suspected factor in 2016 polling error was the failure to weight by education (Kennedy et al. 2016). In the final two weeks of the 2020 election, 317 state-level presidential polls (representing 72% of all polls conducted during this period) provided information on the statistical adjustments accounting for coverage and nonresponse issues; of these 317 polls, 92% accounted for education level in the final results.

- **Polling error was not primarily caused by incorrect assumptions about the composition of the electorate in terms of age, race, ethnicity, gender, or education level**. There is no evidence that polling error was caused by the underrepresentation or overrepresentation of particular demographics. Reweighting survey data to match the actual outcome reveals only minor changes to demographic-based weights.

- **Polling error was not primarily caused by respondents' reluctance to tell interviewers they supported Trump**.[7] The overstatement of Democratic support occurred regardless of mode and the overstatement of Democratic support was larger in races that did not involve Trump (i.e., senatorial and gubernatorial contests).

- **Polling error cannot be explained by error in estimating whether Democratic and Republican respondents voted**. Trump supporters and Biden supporters were equally likely to vote after saying they would. This conclusion is based on validating the vote of registration-based samples shared with the Task Force by some AAPOR Transparency Initiative members.

- **Polling error was not caused by the polls having too few Election Day voters or too many early voters.** Among the 23 state-level presidential polls conducted in the final two weeks that reported how respondents said they would vote, the proportion of Election Day voters closely matched the percentage of certified votes cast on Election Day.

[7] It is plausible that Trump supporters were less likely to participate in polls overall. Nonetheless, *among those who chose to respond to polls* there is no evidence that respondents were lying. A separate and likely problem is that some people chose not to respond to polls at all.

Factors That May Explain the Polling Error

Some explanations of polling error can be ruled out according to the patterns found in the polls, but identifying conclusively why polls overstated the Democratic-Republican margin relative to the certified vote appears to be impossible with the available data. Reliable information is lacking on the demographics, opinions, and vote choice of those not included in polls (either because they were excluded from the sampling frame or else because they chose not to participate), making it impossible to compare voters who responded to polls with voters who did not. Some educated guesses are possible based on patterns emerging from available data but conclusive statements are impossible. It cannot be ruled out that there is a multitude of overlapping explanations for the pattern of polling error.

Voter file information and certified vote information were compared to poll results but the most relevant information is unavailable; for example, it is unknown if Republicans who responded to polls voted differently than those who did not respond.[8] If the voters most supportive of Trump were least likely to participate in polls then the polling error may be explained as follows: Self-identified Republicans who choose to respond to polls are more likely to support Democrats and those who choose not to respond to polls are more likely to support Republicans. Even if the correct percentage of self-identified Republicans were polled, differences in the Republicans who did and did not respond could produce the observed polling error.

This hypothesis is not unreasonable, considering the decreasing trust in institutions and polls especially among Republicans (e.g., Cramer 2016). Trump provided explicit cues to his supporters that polls were "fake" and intended to suppress votes (e.g., Haberman 2020). These statements by Trump could have transformed survey participation into a political act whereby his strongest supporters chose not to respond to polls.[9] If so, self-identified Republican voters who participated in polls may have been more likely to support Democrats than those who chose not to participate in polls. Unfortunately, this hypothesis cannot be directly evaluated without knowing how nonresponders voted. Not only is the percentage of voters who self-identify as Republicans unknown but so too is the vote choice of the self-identified Republicans who chose not to participate.[10]

Many potential explanations for the polling error cannot be evaluated without knowing how respondents and nonrespondents compare. The polls may have differed relative to the 2020 electorate as follows: too many Democrats or too few Republicans, too many new voters or too few, or the wrong percentage of unaffiliated voters.[11] Or perhaps the polling error was caused by differences in the vote choice of the voters who were and were not included in polls (perhaps because the voters refused to participate). Any or all of these possibilities could produce an overestimate of the Democratic-Republican margin, but it is impossible to identify the precise cause(s) of the polling error documented here without knowing the opinions and demographics of voters who were and were not included in polls.

---

[8] An analogous argument extends to unaffiliated voters and voters who were new or newly energized to participate in the 2020 election.

[9] On the importance of elite cues for public opinion and behavior see, for example: Barber and Pope 2018; Bullock 2011; Broockman and Butler 2017; Lenz 2012.

[10] Voter files can be used to estimate partisanship but estimates of individual-level partisanship in the absence of party registration data (or participation in a party primary) are often based on either precinct-level data or an imputation based on demographics (along with correlations between demographics and partisanship among past survey respondents). The former raises questions about the validation of ecological inferences and the latter must assume that the relationship among survey respondents can be used to impute partisanship among survey nonrespondents. But that is precisely the problem of concern.

[11] While we have estimates of these percentages based on voter file records, the characteristics that are of most interest are often estimated.

Even so, the present analyses help quantify the nature of the polling error and suggest what may have happened.

- **At least some of the polling error in 2020 was caused by unit nonresponse**. The overstatement of Democratic support could be attributed to unit nonresponse in several ways: between-party nonresponse, that is, too many Democrats and too few Republicans responding to the polls; within-party nonresponse, that is, differences in the Republicans and Democrats who did and did not respond to polls; or issues related to new voters and unaffiliated voters in terms of size (too many or too few) or representativeness (for example, were the new voters who responded to polls more likely to support Biden than new voters who did not respond to the polls?). Any of these unit nonresponse factors could have contributed to the observed polling error. Without knowing how nonrespondents compare to respondents we cannot conclusively identify the primary source of polling error.

- **Factors that worked well in correcting for nonresponse in previous elections (including demographic composition, partisanship, or 2016 vote) did not render accurate vote estimates for the 2020 election.** Poll data provided by some AAPOR Transparency Initiative members were reweighted to match the 2020 certified outcome. It was necessary to increase the percentage of Republicans (or 2016 Trump voters) and decrease the percentage of Democrats (or 2016 Clinton supporters) in the outcome-reweighted sample. In contrast, there are only slight differences between the originally weighted poll data and the outcome-reweighted data in terms of standard demographic categories.

- **Weighting to a reasonable target for partisanship and past 2016 vote does not fully correct the polling error.** Reweighting the polls to reproduce the 2020 outcome requires a much larger margin for Trump in 2016 than actually occurred among respondents who report voting in 2016. The larger 2016 margin for Trump among those who reported voting for Trump in 2016 could be caused by the following: an issue with the weighting targets, i.e., the implied vote share among 2016 voters who voted in 2020 was different from the 2016 actual outcome; or differences in opinion within groups that responded, e.g., the 2016 Trump supporters who responded to polls were more likely to vote for Democrats than those who did not. It is impossible to know which caused the larger 2016 margin.

- **It is possible that 2020 pre-election polls were not successful in correctly accounting for new voters who participated in the 2020 election.** There were many new voters in 2020 and it is unclear whether the proportion of new voters in the polls matched the proportion of actual new voters. It is also unclear whether the new voters who responded to polls had similar opinions to those who did not respond. Given the relative proportion and self-reported voting behavior of these new voters in the data available to the Task Force, this group of voters pushed the overall polling margins in the Democratic direction. Error in polling this group could have produced the observed polling error.

INTRODUCTION

AAPOR Executive Council charged the Task Force on 2020 Pre-Election Polling with the following.

1. Evaluate the accuracy of 2020 pre-election polling for both the primaries and the general election on the presidential race and other races. To facilitate this analysis, the Task Force should attempt to gather the necessary information for each poll on an ongoing basis during the election season.
2. Where necessary, examine why specific polls and/or methodologies failed to correctly estimate support for major party candidates, whether at the national level or state level.
3. Review variation by different survey methodologies.
4. Identify significant differences between pre-election polls in 2020 and polling in prior election years including 2018, 2016, and 2014.
5. Create an archive of the data collected for the analysis for 2020 and for prior years to enable AAPOR to continue this work in future years.

The Task Force report analyzing the performance of polling in the Democratic presidential primaries was released in October 2020.[12] This report analyzes the performance of pre-election polls leading up to the general election held on November 3, 2020, by focusing on the presidential contest at the national and state levels as well as senatorial and gubernatorial contests. This report provides a comprehensive summary of the overall performance of public polls and it does not evaluate the performance of specific polls or pollsters.

To evaluate the performance of the 2020 general election polls the Task Force collected every public poll in existing databases (e.g., Real Clear Politics, FiveThirtyEight) and actively monitored media stories to collect polls that were not included in existing databases. All told, the topline information on more than 2,800 public polls was collected for analysis. No polls were intentionally excluded from our analysis, but nonpublic polls (e.g., internal candidate and party polls that were not made public) were necessarily excluded. For each poll, we collected all available public material, including news coverage, press releases, cross-tabulations, and methodology reports. Using this information, we classified polls according to their mode and other information relevant for assessing variation in polling error, such as the field period, sample source, the population represented (e.g., registered voter, likely voter), and demographics used in weighting procedures.

We collected information on every public poll conducted during the election cycle we could find, but we followed past practices and focused on the average performance of polls conducted in the final two weeks of the campaign (i.e., polls with a field period ending between October 21 and November 3, 2020). The Task Force is exceptionally grateful for the individual-level data shared with us by AAPOR Transparency Initiative members and the ability to use Task Force members' access to private voter files. These data were essential for our investigations.

[12] https://www.aapor.org/Education-Resources/Reports/2020-Pre-Election-Polls-Performance-of-the-Polls-i.aspx

The report proceeds in two parts. Part one describes the performance of 2020 pre-election polls, including how polling error varies across types of races, modes of interviewing respondents, and states.

Part two begins the process of identifying what did not contribute, and what might have contributed to, the polling error we observed. This analysis is more speculative and limited. We are able to rule out some factors, but our conclusions are limited given the available data. Without knowing how the voters who were included in polls differed from those voters who were not, including how the opinions and vote choice of respondents compared to the opinions and vote choice of nonrespondents, we cannot conclusively determine the source(s) of polling error in 2020. We follow the available data as best we can and we offer some suggestions as to what might have happened, but many explanations could produce the same patterns we document. We do not presume to be able to provide the last word on what happened, but we hope our efforts help guide continued exploration and examination into the extremely difficult task of predicting close elections in a highly partisan and politicized environment.

PART ONE: PRE-ELECTION POLLING IN 2020

Part one describes the types of polls conducted in 2020 and how well they did with respect to polling error, using several measures across the national- and state-level races. Each section begins with the key takeaways for that section. These concise conclusions are followed by a summary of the data and analysis behind them. The polling error is described in eight sections as follows.

- Section 1 describes the number of polls and the types of polling ahead of the 2020 election, including the number of polls conducted at the national and state levels for presidential, gubernatorial, and senatorial contests, and how 2020 polls compare to 2016 polls in terms of how respondents were interviewed.
- Section 2 defines the metrics used to evaluate the accuracy of pre-election polls.
- Section 3 compares the performance of 2020 pre-election polls to the performance of polls in prior election cycles, according to evaluations conducted by AAPOR and the National Council on Public Polls (NCPP).[13]
- Section 4 looks at how well polls were able to identify the winning candidate given the reported margin of error.
- Section 5 evaluates whether or not the performance of polls depends on how respondents were interviewed. While comparisons are difficult given the differences between polls and pollsters, no mode was obviously superior to the others.
- Section 6 shows that the overstatement of the Biden-Trump margin was larger in states that were more supportive of Trump in 2016. This is true even after accounting for between-state differences plausibly related to the difficulty of polling in a state.
- Section 7 shows that polls overstated the Democratic margin for senatorial and gubernatorial candidates more than they overstated Biden-Trump margin within the same state.
- Section 8 compares the polls' average topline support for Biden and Trump to the certified vote share. The polls appear to understate the support for Trump and overstate the support for Biden.

[13] Unlike postmortems that focus on a subset of the polls or the performance of individual pollsters (e.g., Silver 2021), the present analysis includes every public poll.

1. The Extent and Nature of 2020 Pre-Election Polls

Throughout 2020 there were 529 national popular vote polls by 97 pollsters, 1,572 state-level presidential polls by 221 pollsters, and 757 senatorial-gubernatorial poll results by 161 pollsters.

Based on news reports, press releases, and other collections of public poll results, the Task Force collected information on 2,858 polls containing a general-election horserace question. Private polls that were conducted on behalf of candidates, parties, and organizations and not publicly reported were excluded from our analysis. Table 1 reports the number of polls and pollsters by contest and proximity to Election Day.

There were 529 national presidential polls collected over the course of the campaign, including 66 polls conducted by 38 different pollsters in the last two weeks of the campaign, that is, polls with a field period ending between October 21, 2020, and November 3, 2020.

At the state level, the Task Force collected information from 1,572 state-level presidential pre-election polls by 221 different pollsters starting in late March 2020, including 941 polls in states where the final margin in the 2016 presidential election was five percentage points or fewer.

	Number of Polls	Number of Pollsters
National Presidential: All	529	97
National Presidential: Last 2 Weeks	66	38
National Presidential: Last Week	41	31
National Presidential: Last 3 Days	20	18
State-Level Presidential: All	1,572	221
Competitive States Only	941	137
State-Level Presidential: Last 2 Weeks	438	90
Competitive States Only	250	63
State-Level Presidential: Last Week	288	62
Competitive States Only	173	49
State-Level Presidential: Last 3 Days	181	26
Competitive States Only	92	22
Senate-Gubernatorial: All	757	161
Senate-Gubernatorial: Last 2 Weeks	181	66
Senate-Gubernatorial: Last Week	124	44
Senate-Gubernatorial: Last 3 Days	63	15

Table 1. Number of 2020 Public Polls Collected by the Task Force on 2020 Pre-Election Polling by Timing and Races Covered. *Last 2 Weeks* includes any polls with a field period ending on or between 10/21/20 and 11/3/20, *Last Week* includes any poll with a field period ending on or between 10/27/20 and 11/3/20, and *Last 3 Days* includes any poll with a field period ending on or between 10/31/20 and 11/3/20. *Competitive States* are states where the certified 2016 margin was five percentage points or fewer in the presidential contest.

Polling was prominent throughout 2020, although we focus on the "Last 2 Weeks" for continuity with prior AAPOR reports. The prevalence of pre-election polling throughout 2020 is illustrated in Figure 1, which graphs the timing of the 1,572 public state-level presidential polls relative to Election Day. The largest concentration of polls occurred within the last five days of the election, but many polls were conducted and released throughout the year 2020.

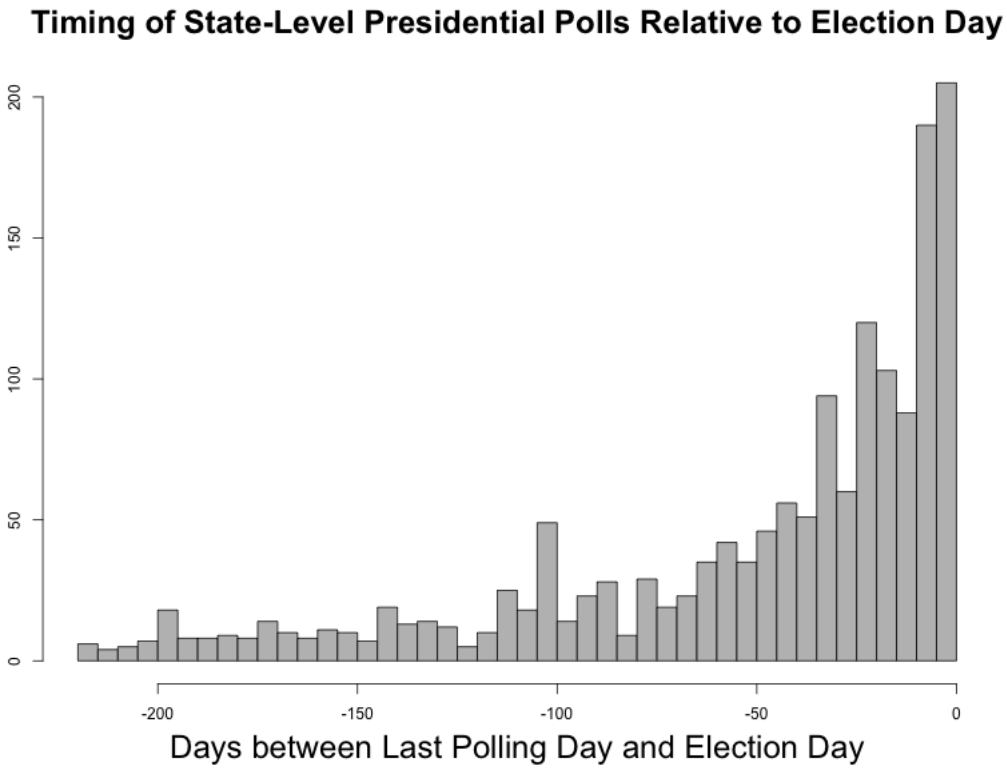


Figure 1. Number of State-Level Public Presidential Polls by Days Before the 2020 Election. This histogram reports the number of polls ending in each five-day period prior to Election Day for all 1,572 state-level public presidential polls that were collected by the Task Force on 2020 Pre-Election Polling. No public poll was intentionally excluded.

1.2 Comparing 2016 and 2020 Survey Modes

Pre-election polls in 2020 were more likely to be done online or using mixed methods than in 2016. More than 60% of all polls were done online and phone polls were less frequently used in 2020 relative to 2016.

To compare 2020 with 2016, each poll was classified by mode of interviewing, according to the categories outlined in the 2016 Task Force report and the descriptions in public releases. Information from pollsters was sometimes missing or difficult to interpret.[14] Nonetheless, the comparisons graphed in Figure 2 reveal several important trends.

First, the use of online polls increased sharply between 2016 and 2020. In 2016, only 38% of national presidential polls were conducted online but that percentage grew to 64% in 2020. The percentage of online polls in competitive battleground states increased from 38% in 2016 to 50% in 2020. Sixty percent of all 2020 state-level presidential polls were conducted online.

Second, the share of phone polls with human interviewers decreased for national polls but remained largely stable among state-level presidential polls. In 2016, 36% of national phone polls used human interviewers and random-digit-dialing (RDD); in 2020, this figure dropped to 6%. The percentage of state-level presidential polls using human interviewers to interview RDD or registration-based (RBS) samples in competitive states remained relatively steady, increasing only slightly from 18% in 2016 to 19% in 2020. [15]

Third, for state-level polls, the percentage relying exclusively on interactive voice response (IVR) technology fell from 14% in 2016 to almost none in 2020. This decline was offset by an increase in polls using multiple interviewing modes.

[14] As emphasized below, considerable variation exists in the methods used within each category. That makes it difficult to interpret between-mode differences. Using publicly available information on the methods of interviewing, the polls were classified as follows: *Live Phone (RDD)* refers to polls conducted by human interviewers calling numbers generated via Random Digit Dialing (RDD); *Live Phone (RBS)* refers to polls conducted by human interviewers calling numbers from a voter registration list (RBS); *Phone & Online* refers to polls that combined interviews conducted by human interviewers, using either RDD or RBS samples, and interviews conducted online; *IVR Only* refers to polls conducted using interactive voice recording technology for telephone numbers obtained using either RDD or RBS samples; *IVR & Phone* refers to polls that combined interviews using interactive voice recording and human interviews for telephone numbers obtained using either RDD or RBS samples; *IVR & Online* refers to polls using interactive voice recordings for telephone numbers obtained using RDD or RBS samples, and polls conducted online; *Online* refers to polls conducted online; and *Other/Unknown* includes polls conducted using relatively rare methods (e.g., Text/SMS) or polls conducted using unreported methods.

[15] Registration Based Sampling (RBS) refers to polls that attempt to contact respondents from a voter registration list that includes contact information. (See, for example, Kennedy et al. 2018). This contact information is not present for all voters on the list; there can be can be errors in matching contact information to voter records; and there can be variation in how the samples are selected, e.g., by quotas based on information in the voter registration list. Such differences between RBS-based polls (relating to lists and sampling methods) are not considered in this overview.

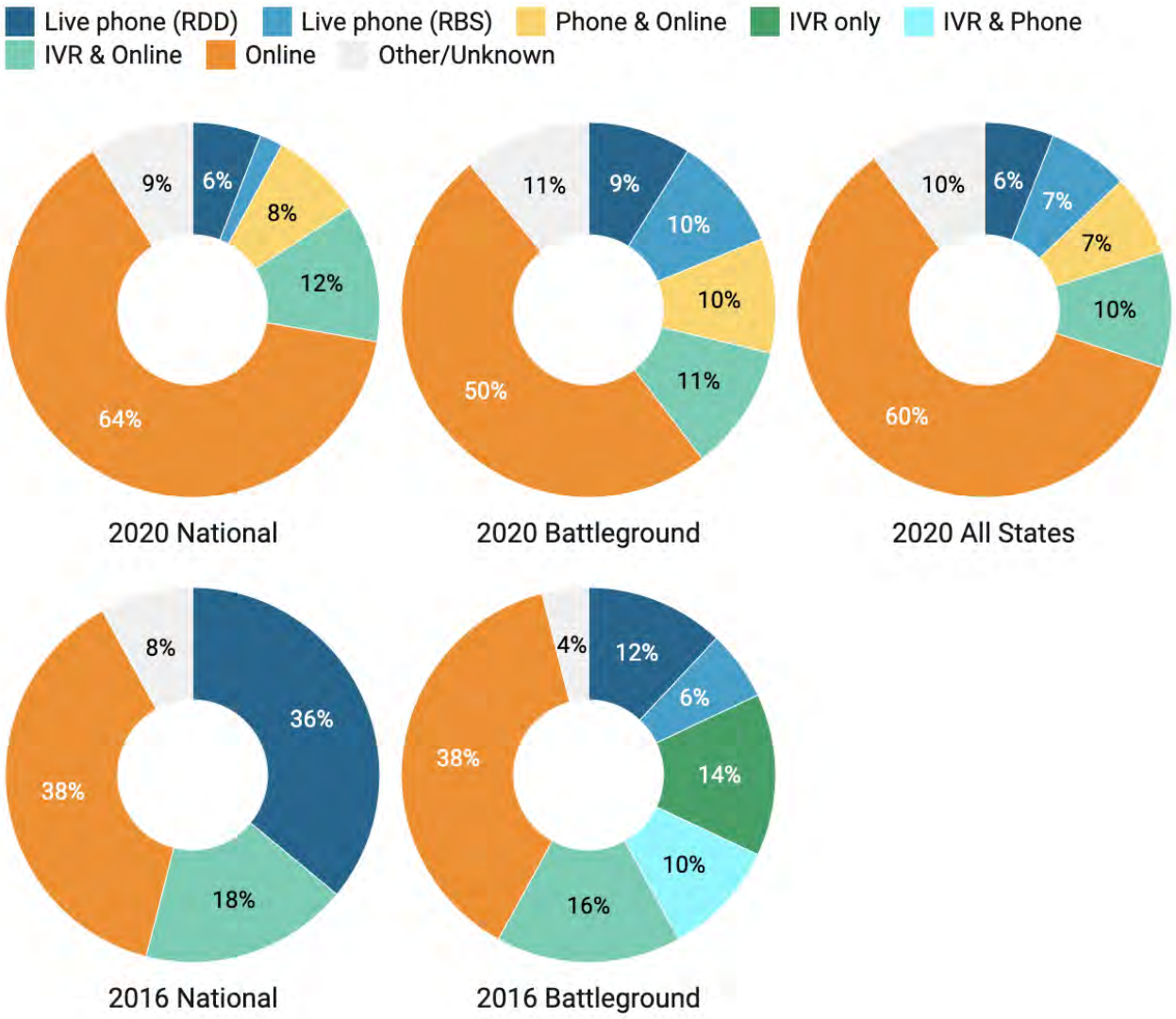


Figure 2. Distribution of Mode of Data Collection for State-Level and National Pre-Election Polls, 2016 and 2020. All polls whose last field period day was between 10/21/20 and 11/3/20 are included. The 2016 results are from Figure 3 of the 2016 AAPOR Task Force report (Kennedy et al.).[16]

Changes in interviewing methods are important not only for describing how polling has changed since 2016 but also for identifying factors contributing to variations in the overall performance of polls varies over time. Over-time differences in performance *could be* due to changes in the types of polls (i.e., how respondents are selected and interviewed and how the results are adjusted to correct for nonresponse and noncoverage) *or* changes in the electoral environment, which may make accurate polling difficult to achieve regardless of how polls are conducted. Changes in the electoral environment may include, for example, third-party candidates or the difficulty of identifying voters. Even though performance can be compared across time with relative ease, it can be very difficult to interpret observed differences.

[16] https://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx

2. Measuring the Accuracy of Pre-Election Polls

There are several ways to analyze the performance of polls. Unlike in 2016, only a small fraction of respondents in 2020 did not choose a major-party candidate when asked.

There are several ways to measure the accuracy of pre-election polls. In prior evaluations, AAPOR Task Forces have used a pair of measures previously employed by the National Council on Public Polls (NCPP): signed error and absolute error.[17] These measures are defined in terms of the 2020 presidential election as follows.

- Absolute error is the absolute value of the margin in the poll (%Biden – %Trump) minus the same margin in the certified vote (%Biden – %Trump).

  |(% Biden Poll – % Trump Poll) – (% Biden Vote – % Trump Vote)|

- Signed error is the same calculation without the absolute value.

  (% Biden Poll – % Trump Poll) – (% Biden Vote – % Trump Vote)

The absolute error compares the margin in the poll to the margin in the certified vote and then takes the absolute value. For example, if Biden beat Trump by seven points in the certified vote, a poll showing a 10-point Biden lead over Trump would produce an absolute error of 3 points, that is, |10 – 7| = |3| = 3. Similarly, if Trump beat Biden by seven points in the certified vote, a poll showing a 10-point Trump lead over Biden would also produce an absolute error of 3 points, that is, |-10 – (-7)| = |-3| = 3.

The signed error describes the direction of error, that is, whether the polling margin overestimates or underestimates the candidate's certified margin of victory. Using the same examples above, if Biden beat Trump by seven points in the certified vote and the poll had a 10-point Biden lead then the poll's signed error would be +3, that is, 10 – 7 = 3. However, as another example, if Trump beat Biden by 10 points and the poll had a only a seven-point Trump lead then the signed error would be -3, that is, 7 – 10 = -3.

In this analysis, positive signed errors indicate polling margins that are too large for Biden (or another Democratic candidate) relative to the certified vote margin (i.e., an overstatement of the Democrat minus Republican margin). Negative signed errors indicate polling margins are too large for Trump (or another Republican) relative to the certified vote margin (i.e., an understatement of the Democrat minus Republican margin).

Considering both absolute and signed errors is important because positive signed errors (poll margin too Democratic) and negative signed errors (poll margin too Republican) offset when averaging signed errors across polls. A poll that is 10 points too favorable towards Biden on the margin (+10) and a poll that is 10 points too favorable towards Trump on the margin (-10) average out to zero points of signed error. In contrast, the average absolute error on the margin would be 10 points because the absolute error quantifies the size of the polling error on the margin regardless of its direction.

[17] The website of the National Council of Public Polls is available here, although it is no longer being updated: http://www.ncpp.org/

When computing the absolute and signed errors we follow past practices and use the topline candidate marginals reported by each poll. As an example, if a poll reported that 49% backed Biden, 45% backed Trump, and the remaining 6% chose an alternative response (e.g., "other" or "don't know"), then the values of 49% and 45% are used to compute the poll's margin (here 49 – 45 = 4) that is compared against the certified vote margin.

The proportion of undecided voters was relatively small in 2020, unlike in 2016 when a sizable proportion of respondents were undecided (e.g., Silver 2017). In 2020, respondents rarely ever gave an answer other than Biden or Trump when asked. As Table 2 summarizes, only 4% responded with something other than Biden or Trump when asked in state-level presidential polls in the last two weeks. Respondents were also often willing and able to choose a candidate in polls for senatorial or gubernatorial contests.

	Number of Polls	Mean % NOT Choosing a Democrat or Republican
State-Level Presidential: All	1,572	5.9
State-Level Presidential: Last 2 Weeks	438	4.0
State-Level Presidential: Last Week	288	4.0
State-Level Presidential: Last 3 Days	181	3.5
Senate-Gubernatorial: All	757	10.0
Senate-Gubernatorial: Last 2 Weeks	181	6.1
Senate-Gubernatorial: Last Week	124	5.8
Senate-Gubernatorial: Last 3 Days	63	4.8

Table 2. Mean Percentage of Respondents Choosing a Response Other than a Major Party Candidate. This percentage was computed as follows: 100% – Dem % – Rep %.

3. How Well Did the Polls in 2020 Fare Relative to Prior Years?

On average, national presidential polls had their worst performance in 40 years and state-level presidential polls had their worst overall performance in 20 years. The polls overstated the Biden-Trump margin by 3.9 points in the national popular vote and by 4.3 points in state-level presidential polls.

We begin by considering how well the 2020 pre-election polls performed relative to the pre-election polls of prior election cycles using the traditional measures of polling performance: the average signed error and average absolute error.

To put the polling error of 2020 in a larger context, Figure 3 compares the national 2020 error to the polling error for every presidential election following the infamous 1936 election. Larger polling error has certainly occurred in the past; for example, the 6.0 points of absolute and signed error in 1980. Nonetheless, the average performance of polls in 2020 was among the worst in recent memory. The polling error was similar in 2020 and 1996 but expectations were not high at the time and the 1996 performance was considered acceptable (Mitofsky 1998).[18]

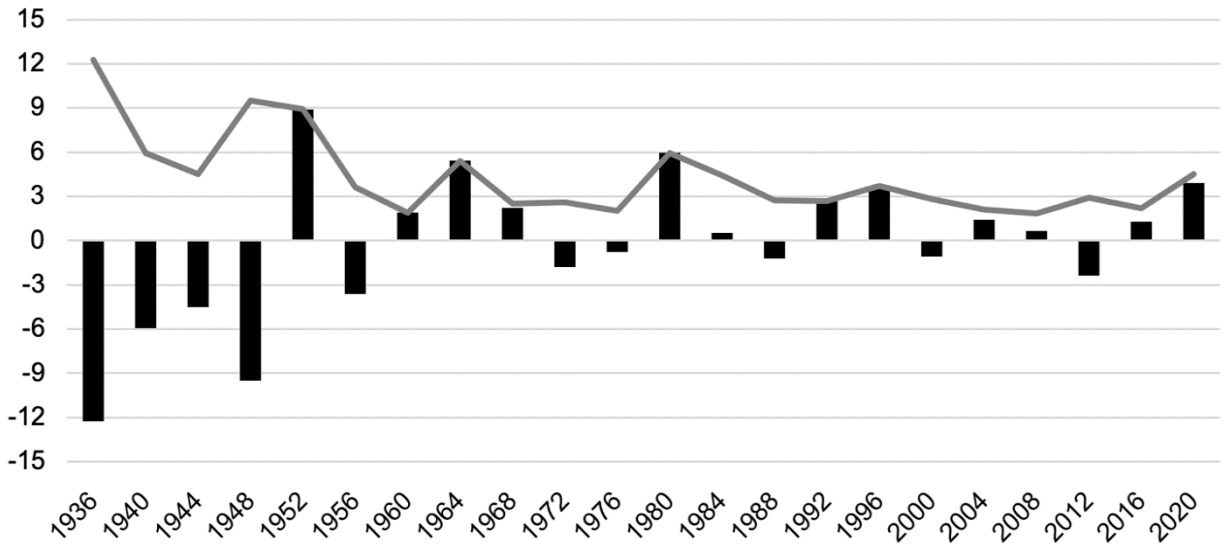


Figure 3. Mean Signed and Absolute Polling Error Over Time: National Presidential Polls. The grey line represents average absolute error on the margin. Bars represent average signed error on the margin. (Positive values indicate overestimation of the Democratic vote margin; negative values indicate overestimation of Republican vote margin.)

[18] Of course, the 1996 election was complicated by Ross Perot, the Reform Party candidate whose presence could have changed the electorate and the difficulty of polling.

Table 3 compares polling error for the six most recent presidential elections. The 2020 pre-election polls that focused on national-level presidential popular vote had an average (mean) signed error of 3.9 percentage points (i.e., they overestimated Biden's margin by 3.9 percentage points). Even among polls conducted within the last week of the election (i.e., polls whose last day in the field was between October 27 and November 3), the average signed error was considerable. The average (mean) absolute error of 4.5 percentage points in the 2020 national polls more than doubled the corresponding absolute error of 2.2 in 2016.[19]

	Mean National Popular Vote Polling Error		Mean State-Level Presidential Polling Error	
Year	Signed Error	Absolute Error	Signed Error	Absolute Error
2000	-1.1	2.8	-1.9	4.6
2004	1.4	2.1	0.8	3.2
2008	0.7	1.8	-0.1	3.4
2012	-2.4	2.9	-2.3	3.5
2016	1.3	2.2	3.0	5.1
2020	3.9	4.5	4.3	5.1
2020 (Last Week)	4.2	4.6	4.1	4.9
2020 (Last 3 Days)	4.4	4.5	3.9	4.7

Table 3. Mean Signed and Absolute Errors for National and State-Level Presidential Polls for Presidential Elections Since 2000. Only polls with a field period ending between 10/21/20 and 11/3/20 are included in the 2020 calculations. *Last Week* results include polls with a field period ending on or between 10/27/20 and 11/3/20. *Last 3 Days* results include polls with a field period ending on or between 10/31/20 and 11/3/20.

[19] The size of average (mean) polling error is instructive relative to Table 3. The average signed error on the margin (4.3) in the state polls exceeds the percentage of respondents who failed to give a candidate response (4.0); hence, the polling error in 2020 is not explained by the last-minute decisions of undecided voters. As will be shown in part two, there simply were too few late-deciding voters to attribute the observed polling error to them.

State-level presidential polls performed similarly. Among all polls conducted in the final two weeks, state-level polls overestimated Biden's margin by 4.3 percentage points, the highest signed error in at least 20 years. The overall average absolute polling error was 5.1 percentage points, the same average error as 2016. Calculating medians instead of averages to minimize the impact of outliers changes the characterization only slightly: The median signed polling error was 4.0 percentage points and the median absolute polling error was 5.0 percentage points.

Moreover, polls conducted in the final week did about as well as those done within the last two weeks. In terms of signed error, state polls conducted in the last week overstated Biden's margin by 4.1 percentage points, and polls conducted in the final three days overstated Biden's margin by 3.9 percentage points. While the size of the average signed error decreased among polls done closer to Election Day in 2020, the signed error of 3.9 points for polls done in the last three days of the 2020 election was still larger than the average signed error for polls done in the last two weeks of the 2016 election (+3.0).

To facilitate comparisons, Figure 4 graphs the average signed and absolute error in state-level presidential polls over time to highlight that although the average absolute error was similarly high in 2016 (and nearly as high in 2000), the 4.3 points of signed error was the highest in recent years. Unlike in earlier election cycles, signed and absolute errors are similar in 2020, which is as expected for errors largely defined by the overstatement of the Democratic-Republican margin.

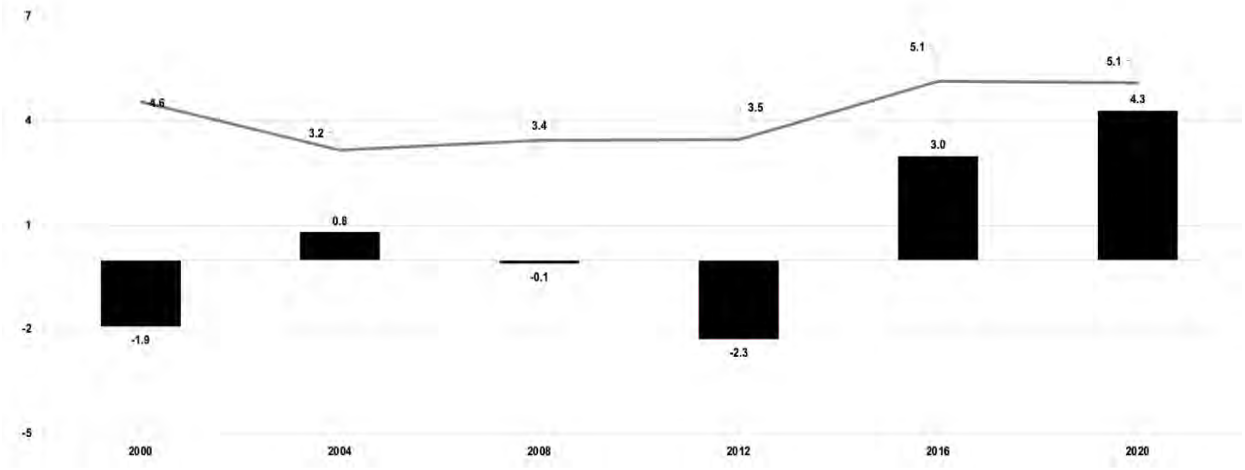


Figure 4. Mean Signed and Absolute Polling Error Over Time for State-Level Presidential Polls. The grey line represents average absolute error on the margin. Bars represent average signed error on the margin (positive values indicate overestimation of the Democratic vote margin; negative values indicate overestimation of Republican vote margin).

To illustrate the stability of the polling error in state-level presidential polls across the election campaign, Figure 5 graphs the distribution of signed error and absolute error for all state-level presidential polls conducted in the last two weeks, the last week, and the last three days of the 2020 election. Regardless of the timeframe, the distributions and the averages (indicated by a vertical line in each histogram) of the polling error are nearly identical.[20]

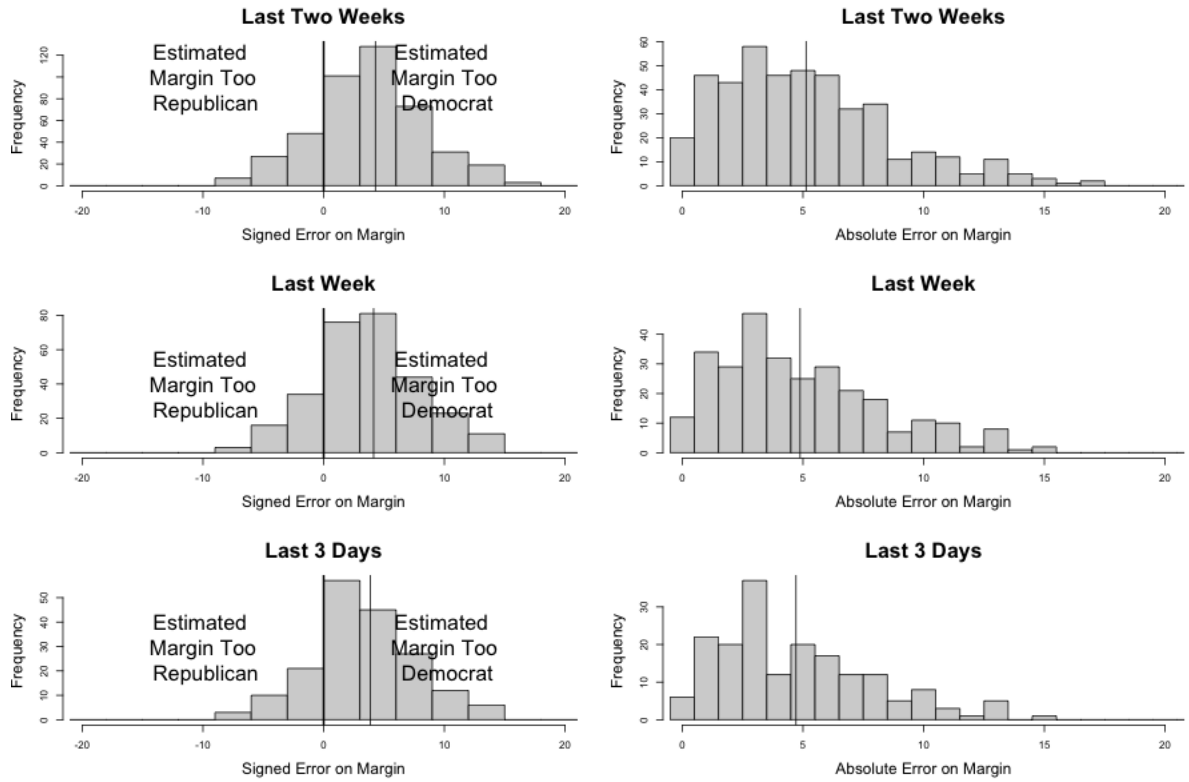


Figure 5. Distribution of Signed and Absolute Error in State-Level Presidential Polls. *Last Two Weeks, Last Week,* and *Last 3 Days* indicate field periods ending between 10/21/20 and 11/3/20, between 10/27/20 and 11/3/20, and between 10/31/20 and 11/3/20, respectively. Vertical lines denote the average of each distribution.

3.1 Senatorial and Gubernatorial Polls

> *Polls of senatorial and gubernatorial contests overstated the Democratic-Republican margin by 6.0 points relative to the certified vote, on average. The six-point overstatement of the margin for Democratic candidates was larger than the polls' overstatement of the Biden-Trump margin. The average absolute error of 6.7 points was the largest in at least 20 years.*

Polls for senatorial and gubernatorial elections were examined to determine whether the polling error in 2020 was confined to polls asking about support for Trump.

[20] Figure B1 in Appendix B compares the distribution of signed error for polls where more than 5% of the sample chose neither Biden nor Trump to the distribution of signed error among polls with fewer respondents doing so to determine how the presence of more undecided voters in earlier polls affects the signed error. Replicating the analysis using polls with relatively few undecided voters reveals similar conclusions.

Table 4 compares the performance of polls in 2020 to the performance in earlier elections according to prior evaluations.[21] In some cases, the sources included state-level presidential polls; in other cases, the sources excluded state-level presidential polls. In these results, both types are averaged for the sake of continuity and comparison.

Year	Election Day	Polls Included (Source)	Mean Absolute Error	Mean Signed Error
2002	11/5/2002	Senate, Governor (NCPP)	4.8	NA
2004	11/2/2004	President, Senate, Governor (NCPP)	3.4	NA
2006	11/7/2006	Senate, Governor (NCPP)	4.0	NA
2008	11/4/2008	President, Senate, Governor (NCPP)	4.0	NA
2010	11/2/2010	Senate, Governor (NCPP)	4.8	NA
2012	11/6/2012	President, Senate, Governor (NCPP)	3.8	NA
2014	11/4/2014	Senate, Governor (NCPP)	5.8	4.1
2018	11/6/2018	Senate, Governor (AAPOR)	4.6	0.1
2020	11/3/2020	Senate, Governor	6.7	6.0
2020	11/3/2020	President, Senate, Governor	6.2	5.5

Table 4. Comparing Gubernatorial and Senatorial Poll Error Over Time. Only polls with a field period ending between 10/21/20 and 11/3/20 are included in the 2020 analysis. This period included 157 senatorial polls and 24 gubernatorial polls. AAPOR did not evaluate the performance of senatorial and gubernatorial polls in 2016.

[21] It can be difficult to interpret the variation in the performance of senatorial and gubernatorial polls over time because the states holding senatorial and gubernatorial contests vary across time. Moreover, unlike presidential contests where the same major-party candidates are on every ballot, the variation in who is running may also impact the difficulty of polling (perhaps by affecting voter turnout).

In terms of average absolute error for the senatorial and gubernatorial polls, Table 4 shows an error of 6.7 percentage points, i.e., the largest in 20 years regardless of whether or not presidential state-level polls are included. It is unclear how this compares to the error in 2016 because the 2016 AAPOR Task Force did not examine the performance of senatorial and gubernatorial polls, but 6.7 points in 2020 is nearly 1.5 times the average absolute error of 4.6 points in 2018. This increase is not driven by a few extreme errors since the median is nearly the same as the average.

The average absolute error on the margin slightly decreases to 6.2 percentage points when including presidential polls, indicating larger polling errors for senatorial and gubernatorial contests than for the presidential contest.

In terms of the signed error for senatorial and gubernatorial contests, polls overstated the Democratic-Republican margin by six points regardless of whether we use the mean or median to summarize the performance of polls. The average signed error decreases from 6.0 to 5.5 points when presidential polls at the state level are included, indicating that the Democratic margin was overstated more in senatorial and gubernatorial contests than in presidential contests.

Figure 6 graphs the distributions of polling error for gubernatorial and senatorial polls in a manner analogous to Figure 5, which shows comparable distributions for presidential polls at the state level. The frequency (vertical axis) of the error (horizontal axis) is presented for three time periods with bottom histograms closest to Election Day. Distributions of signed and absolute errors are plotted separately in histograms on the left and right, respectively.

Analogous with the state-level presidential poll distributions, most senatorial and gubernatorial polls overstated Democratic margins relative to the certified vote and this did not change closer to Election Day. The polling error for polls completed in the final three days was very similar to the polling error for polls completed in the final two weeks.

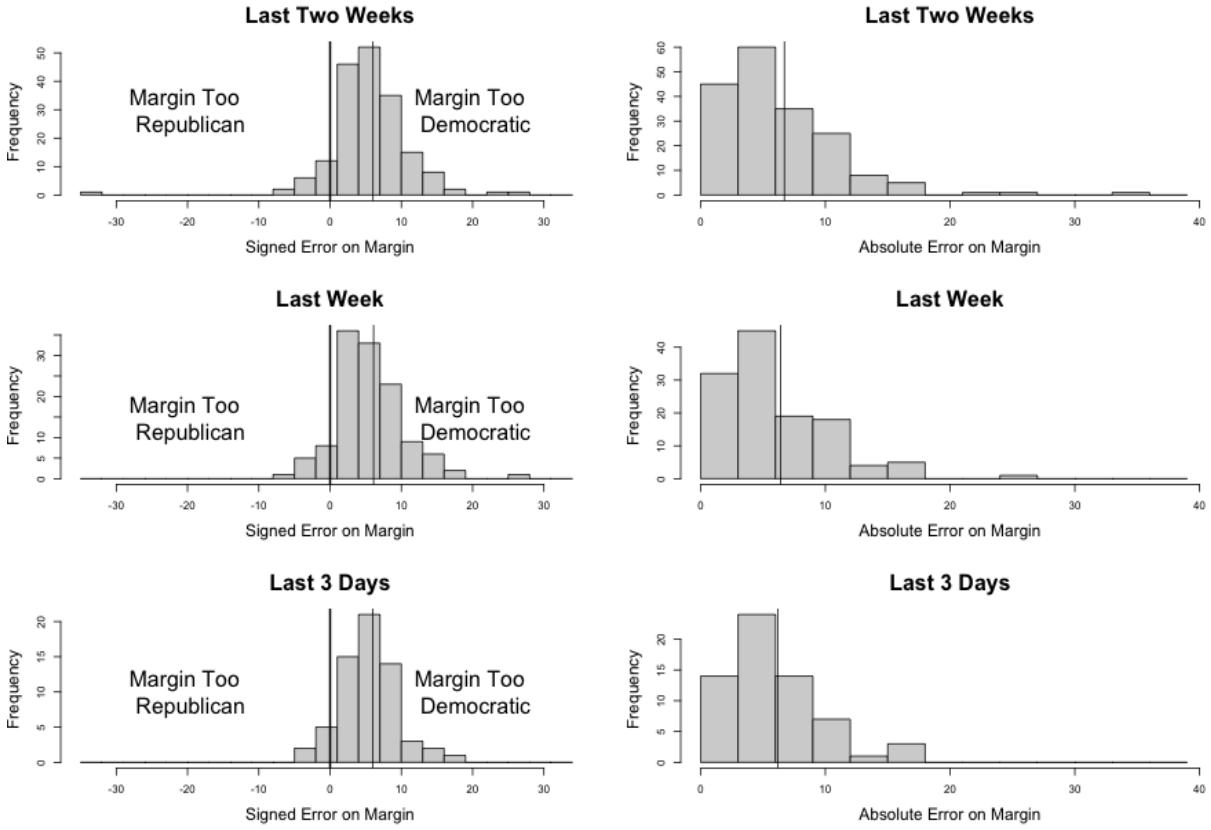


Figure 6. Distribution of Signed and Absolute Error in Gubernatorial and Senatorial Polls. *Last Two Weeks* includes polls with a field period ending between 10/21/20 and 11/3/20, *Last Week* includes polls with a field period ending between 10/27/20 and 11/3/20, and *Last 3 Days* includes polls with a field period ending on or between 10/31/20 and 11/3/20. Vertical lines denote the average of each distribution. The height of the histogram indicates the number of polls within each three-point range of error.

4. Identifying the Winning Candidate in 2020

Seventy-nine percent of state-level presidential polls and 69% of senatorial and gubernatorial polls had the winning candidate in the lead. Among polls with a candidate margin more than twice the reported margin of error, 98% of state-level presidential polls and 90% of senatorial-gubernatorial polls correctly identified the winning candidate. Accounting for the uncertainty of polling is essential to contextualize the results in close contests. Polls are often unable to distinguish the margin from a tie.

How well polls identified the winning candidate was evaluated using several measures to account for narrow margins both in the polls and in the outcomes. Regardless of the closeness of the poll or the election, one simple approach is to note whether or not the candidate leading in the poll won the election.

Two additional measures were applied to evaluate polls that reported a "margin of error." Out of the 438 state-level presidential polls conducted in the last two weeks, there were 348 (or 79%) that reported a margin of error along with the poll margin.[22]

The second measure only counts polls as correct if the winning candidate had a *poll margin* (also called candidate margin) greater than twice the margin of error. This size of poll margin is statistically distinguishable from zero at conventional levels. In other words, among polls that reported a candidate margin greater than twice the margin of error, how likely was it that the leading candidate won? Although it is not the statistically correct calculation to quantify the standard error associated with the difference between two correlated proportions from a multinomial distribution, it is easy to place an upper bound on the precision if the margin is "within the margin of error" using reported information.[23] This measure effectively removes polls with narrow margins relative to the margin of error.

"Twice the margin of error" is much more conservative than the actual standard for determining whether a poll margin is statistically distinguishable from zero using conventional levels of statistical significance. A conventional statistical test is a 95% confidence level but this "twice the margin of error" approach is a reasonable choice, especially considering the many sources of error that go beyond sampling variability (Goel and Rothschild 2016).[24] (Table B1 in Appendix B replicates the analysis using just the margin of error and makes the point that coverage focusing on the margin of error for an estimated proportion rather than the margin of error for the difference in proportions creates a mistaken impression that polls are more precise than they actually are.)

[22] Some online polls report a "Bayesian credible interval" or another measure of uncertainty; that said, 74% of the online polls reported a quantity called a "margin of error." Only measures called a "margin of error" are included in this evaluation. The results of this evaluation could be replicated using the sample size of each poll to calculate the implied margin of error for every public poll, but the qualitative conclusions would be unchanged compared to the conclusions reached using only polls that report a "margin of error" in the public material.

[23] If X is the estimated support for candidate 1 and Y is the estimated support for candidate 2, the variance of the margin of the poll is the difference between the two random variables X and Y reduced by twice the covariance: Var(X-Y) = Var(X) + Var(Y) – 2COV(X,Y). Given that support for Biden and Trump from the same survey are obviously correlated, using the twice the margin of error is higher than the actual standard error on the difference (Kish 1968, Scott and Seber 1983, Franklin 2007).

[24] Technically, whether or not a margin is statistically distinguishable from zero would require using the standard error on the difference of the multinomial distribution. We have focused on a quantity that is more easily accessible to explain and analyze, i.e., twice the reported margin of error.

The third measure focuses on the *certified margin* (or vote margin) being twice as large as the reported margin of error. In these polls, it should be possible to identify the winner from the size of the electoral victory and the margin of error. This measure differs from the second because it only considers the accuracy of polls in elections with a certified outcome that was "not close" (defined as a vote margin at least twice as large as the margin of error for each poll). Again, what fraction of polls correctly identify the eventual winner when the margin is statistically distinguishable from zero according to the reported margin of error? This measure effectively removes polls from closely contested states (relative to the margin of error).

Type of Poll	Percent Identified Winning Candidate: Simple	Percent Identified Winning Candidate: Poll Margin > 2xMoE	Percent Identified Winning Candidate: Vote Margin > 2xMoE
National-Level Presidential	100 (N=66)	100 (N=66)	100 (N=66)
State-Level Presidential	76 (N=438)	98 (N=223)	94 (N=215)
State-Level Presidential in a state with Gubernatorial or Senatorial Contest	79 (N=261)	100 (N=86)	96 (N=89)
Gubernatorial and Senatorial	69 (N=181)	90 (N=69)	78 (N=35)
Senatorial	66 (N=157)	87 (N=54)	77 (N=56)
Gubernatorial	72 (N=24)	100 (N=15)	89 (N=9)

Table 5. Percent of Polls that Correctly Identified Election Winner from Polls Conducted in the Last Two Weeks by Contest Type. The "Simple" estimate is based on whether or not the winning candidate was in the lead regardless of the margin of error. The second measure reports on the percentage of polls with the winning candidate leading among polls with a candidate margin at least twice as large as the poll's reported margin of error. (The average reported margin of error was 3.9.) The third measure reports the percentage of polls with the winning candidate leading when the final certified margin is at least twice as large as the poll's reported margin of error.

All 66 national presidential polls conducted in the last two weeks correctly identified Biden as winning the national popular vote. Among the 438 state-level presidential polls, 76% had the winning candidate ahead. After removing the polls with margins that are within twice the margin of error, 98% of the remaining 223 polls had the winning candidate in the lead. In states where the final margin was twice as large as a poll's margin of error, 94% of the 215 polls correctly identified the winner. The ability of polls to correctly identify the winning candidate when the margin or outcome could be statistically distinguished from zero highlights the importance of accounting for the uncertainty of polls when discussing their results. Losing candidates were more likely to be leading in polls where the candidate margins were not statistically distinguishable from a tie using conventional significance levels.

Even though the "not close" polls performed as expected and accounted for the margin of error in comparison with the candidate margins, most of the politically interesting contests are excluded by such a standard. Only 223 out of the 348 state-level polls (64%) had a candidate margin more than twice the margin of error, meaning the candidate margin could not be statistically distinguished from a tie using standard levels of statistical significance in 36% of the state-level presidential polls.

The performance of senatorial polls was notably worse.[25] Overall, only 66% of the 157 senatorial polls correctly identified the winning candidate. Among the 69 polls that had one candidate leading by more than twice the margin of error, the leading candidate won only 90% of the time. Moreover, in states where the election margin was greater than twice the margin of error in the poll, and hence the polls should have been able to identify the winning candidate, only 78% of the polls identified the winner. Polls were better able to identify the winners in gubernatorial contests, but there were only a few such polls conducted in the last two weeks of the election.

Clearly, knowing the margin of error is essential to confidently identify the winning candidate. Additionally, polling error can be compared with the margin of error to evaluate performance. Did the reported absolute error on the margin contain more or less error than expected from the margin of error? If the margin of error gives an accurate estimate of uncertainty then most polling errors should be less than twice the reported margin of error. Whether or not the polling error agrees with the reported margin of error provides a benchmark for performance.

Figure 7 plots the distribution of absolute error along with the average margin of error and twice the average margin of error for the 348 (out of 438) state-level presidential polls that reported a margin of error. The overall average margin of error was 3.9 points.

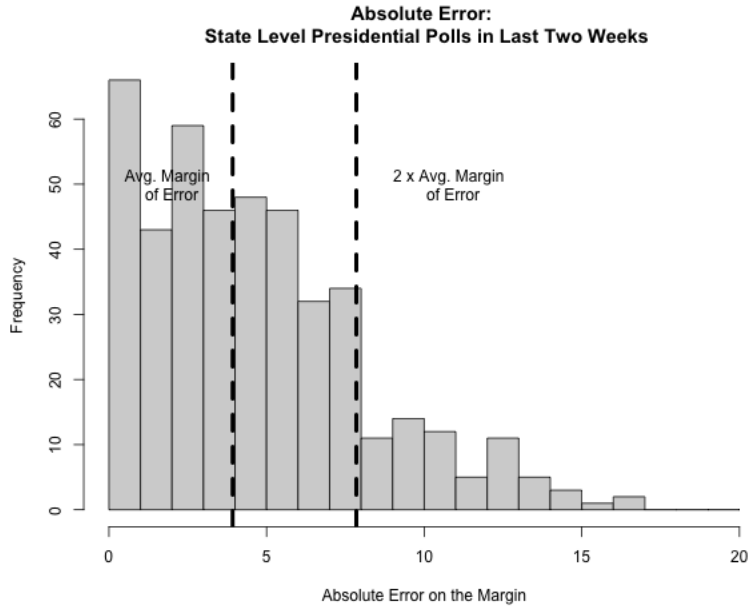


Figure 7. Comparing the Distribution of Absolute Polling Error to the Average Margin of Error Reported by the Polls. Only state-level presidential polls conducted in the *Last Two Weeks* are included.

[25] For the two Georgia senatorial races, the certified vote and outcome of the November general election was used because that election was relevant to the polls conducted for those two races.

Figure 7 reveals that 45% of the state-level presidential polls had an absolute error less than the average margin of error; and 79% had an absolute error less than twice the average margin of error.

A more important finding is that 21% of the polls had an absolute error on the margin that was more than twice the margin of error, i.e., larger than a difference that would be statistically distinguishable from zero. That means the size of the polling error in 21% of the polls was larger than what would be expected based on sampling error alone using a larger-than-usual confidence interval (i.e., twice the margin of error). This finding confirms that many decisions made in conducting pre-election polling can augment the polling error. The margin of error cannot adequately account for all of those decisions.

Comparing the margin of error and the average absolute error also clearly reveals the inadequacy of using the margin of error to describe the error on the margin. In that case, 55% of the polls had an absolute error that was larger than the reported margin of error.

Focusing on the margin of error (or even twice the margin of error) provides an overly optimistic account of the size of the polling error in 2020.

5. Average Performance of 2020 Polls by Survey Mode

> *The Biden-Trump margin was overstated regardless of mode, and no mode was obviously more accurate. Polling error in 2020 for particular methods of interviewing respondents was examined. The polling error was similar despite variation in methods, questions, and the adjustments that were made by pollsters.*

Examining the effects of the mode of interviewing on polling error is useful for understanding the extent of the error but it is difficult to disentangle interview-mode effects from effects that are related to sampling frame coverage, methods of selecting respondents, likely-voter modeling, and the statistical adjustments that are used to correct for nonresponse and noncoverage. (See, for example, Crespi 1988; Cohn 2016; and Groves et al. 2009).

Comparing the performance of polls across modes can be difficult if certain modes are used more frequently in some contests than others. The level of interest in the contest and the costs of polling (e.g., phone polls may not be used as much as online polls in less competitive contests) may complicate comparisons by introducing additional sources of error.

In the analysis of the performance of the polls by interview mode, the focus was primarily on state-level presidential polls, following the precedent of the 2016 Task Force report. State-level presidential polls are the appropriate focus as they ask about the same candidates, which obviates examining whether candidate differences contribute to mode differences.

Among presidential state-level polls it was difficult to determine the methodology of about 10% of the polls (Figure 2), especially for polls conducted by nonmembers of the AAPOR Transparency Initiative. When multiple methods were used for reaching and interviewing respondents, the relative mix was often unclear even when the reported combination was similar.

Another difficulty in comparing the effects of modes on polling error is that some modes were dominated by a single pollster. Since nearly 50% of the *Phone/Online* polls were conducted by the same pollster, it is not clear whether the performance is due to the *Phone/Online* mode or decisions on weighting made by this one pollster.

Figure 8 plots the distributions of absolute and signed error for state-level presidential polls by interviewing mode. (Appendix C repeats the analysis for "mixed" vs. online vs. phone and yields similar conclusions when comparing mixed vs. single-mode surveys.) The violin plots show the distribution of the error for each polling mode, the location and range of the 25th-75th percentiles of the error (the thick black line), the median error (open circle), and the range of the error within 1.5 times the interquartile range (thin black line).[26]

In such distribution plots, the width of the violin-shaped plot reflects the number of polls sharing a particular range of polling error, e.g., among online polls there are many more polls with an absolute error of +5 than there are with an absolute error of +15 and there are no IVR/Online polls with an absolute error of +15. The top and bottom graphs show the distribution of absolute and signed error, respectively. Since the average polling error can differ from the median polling error because of outliers, the average error for each mode is also reported in Figure 8.

[26] Polls were omitted if the mode was unknown, relatively unique (e.g., there were only four IVR-only polls could be identified), or ambiguous (e.g., "email").

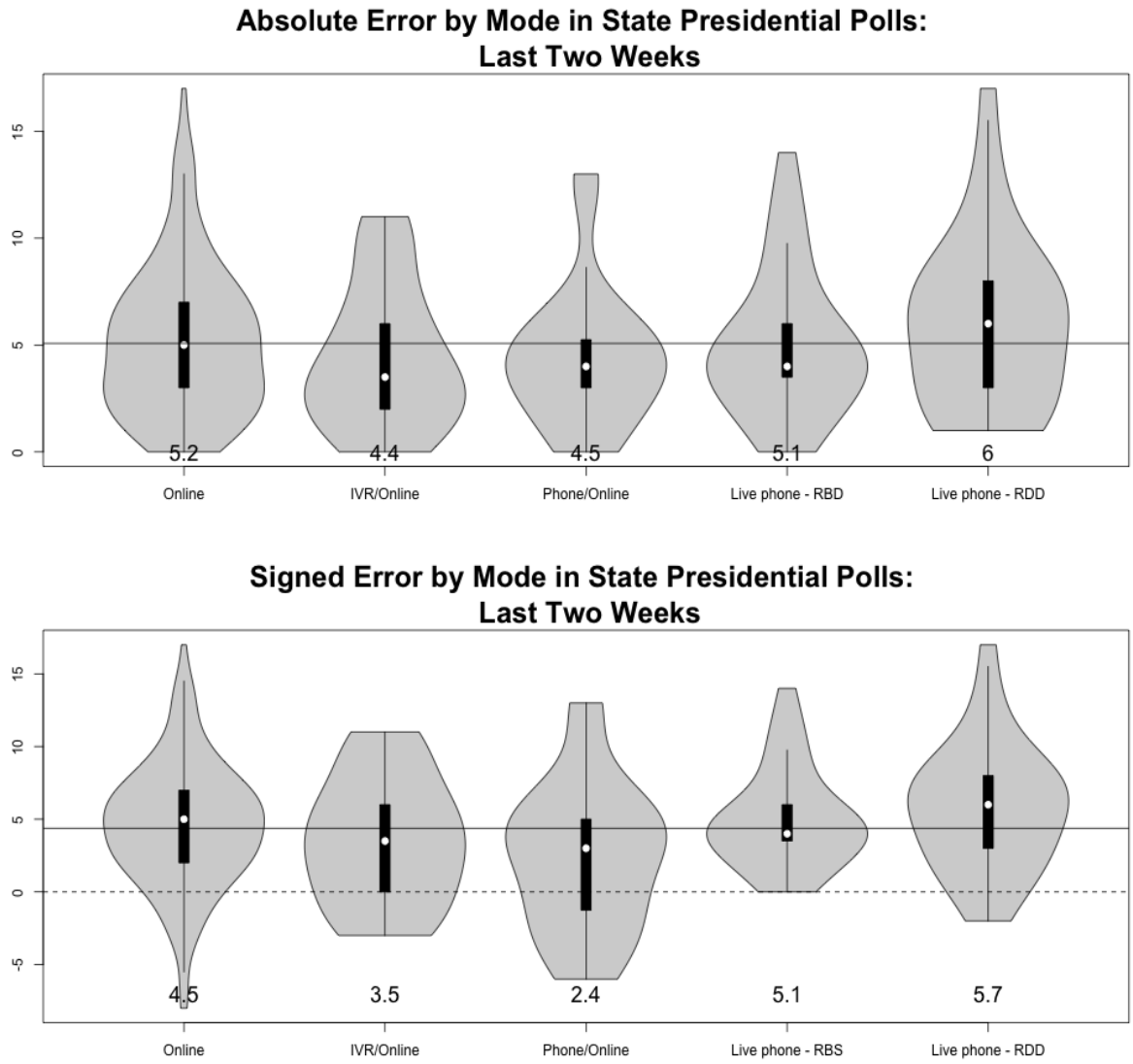


Figure 8. Distributions of Absolute and Signed Error in State Presidential Polls by Survey Mode. The analyses include 262 *Online* polls, 42 *IVR/Online* polls, 32 *Online/Phone* polls, 27 *Live phone - RDD* polls, and 31 *Live phone - RBS* polls. The width of each "violin" reports the distribution of polling error for each mode. The single horizontal line represents the overall error averaged across all modes, while the average polling error for each mode is reported numerically along the bottom.

As the top graph of Figure 8 shows, the polls have similar levels of average absolute error regardless of how respondents were interviewed. The average absolute error varied from 4.4 points for *IVR/Online* polls to 6.0 points for *Live phone – RBS* polls. Conducting a t-test to determine if this average difference is statistically distinguishable from zero reveals that the difference is not distinguishable from zero at conventional levels (p=0.1).

Examining the variation in signed polling error by interview mode reveals a similar conclusion. The average signed error ranged from 2.4 points in *Phone/Online* polls to 6.0 points for *Live phone – RDD* polls.[27] With the exception of *Phone/Online* polls, more than 75% of the polls had a positive signed error, meaning that most polls in most modes overstated the Biden-Trump margin.

Examining the percentage of polls that correctly identified the winning candidate according to the certified vote, *Online* polls did the best (82%) and *Phone/Online* polls did the worst (56%). That *Phone/Online* polls had the lowest average signed error and the highest percentage of losing candidates ahead in their polls highlights the complexity of interpreting polling performances. This seemingly odd pattern of performance is because *Phone/Online* polls, i.e., the polls dominated by a single pollster, were more likely than other polls using other modes to indicate that Trump would win. These polls were also more likely to be conducted in states with narrower margins in 2016 (average margin of 2.4 points). As a result, it may have been both easier to estimate the margin (which was expected to be close) and harder to predict the winner. That is exactly the pattern that is observed.[28]

In Appendix C, Table C1 excludes respondents without an opinion thereby removing question-wording differences; Table C2 focuses only on polls conducted in competitive states, and looks at mixed-mode surveys versus those using a single method of interviewing. The conclusions are similar: The overstatement of Democratic support occurred regardless of the mode of interviewing, and no mode of interviewing was clearly more accurate.

6. Performance of 2020 Polls by Electoral Competitiveness

Polls overstated the Biden-Trump margin more in states that Trump won by larger margins in 2016.

Was polling error largest in states that often vote Democratic, states that often vote Republican, or in closely contested states? Put differently, were the polls more likely to overstate the Democratic-Republican margin in states with more Republican or more Democratic voters? Determining where larger polling error occurred may help narrow the search for possible explanations.

Figure 9 graphs the distribution of absolute (left) and signed (right) error for polls conducted in states Clinton won by more than five points in 2016 (Solid Dem), states where the certified vote difference between Clinton and Trump was five points or fewer in 2016, and states Trump won by more than five points in 2016 (Solid Rep).

[27] This difference is statistically distinguishable from zero but we would urge caution when interpreting this number, as nearly 50% of the polls were from a single pollster and it is impossible to know whether the difference is due to the mode of interviewing or to the adjustments that were being done. Moreover, as we note below, the low signed error for *Phone/Online* polls is offset by a lower ability to identify winners.

[28] By way of contrast, the average 2016 margin for *Online polls* was four points, the average 2016 margin for *IVR/Online* polls was 3.7 points, the average 2016 margin for *Live phone – RBS* polls was 3.1, and the average 2016 margin for *Live phone – RBS* polls was 2.7 points.

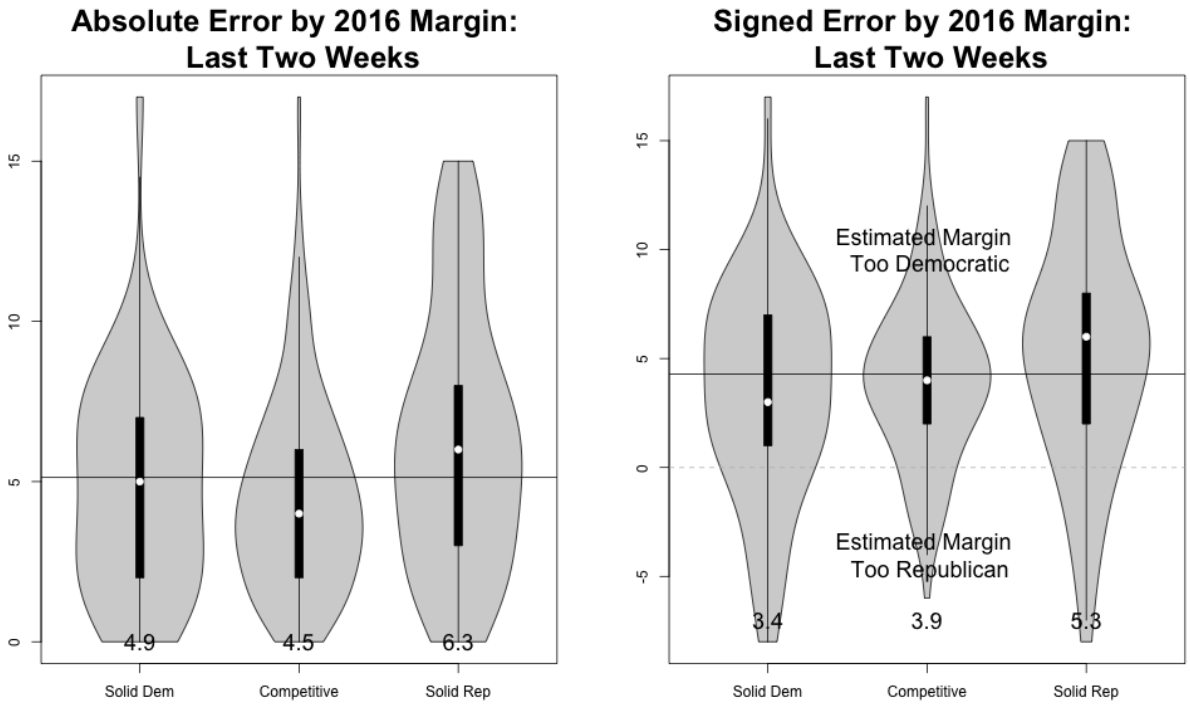


Figure 9. Distribution of Signed and Absolute Error in State-Level Presidential Polls by 2016 State Competitiveness. There are 49 polls in solid Democratic states (won by Clinton in 2016 by at least five points), 239 polls in competitive states (certified vote margin between Trump and Clinton five points or fewer in 2016), and 139 polls in solid Republican states (won by Trump in 2016 by at least five points). Horizontal lines denote the overall average polling error.

The left-hand side of Figure 9 shows that the average absolute error was lowest in states where the final 2016 margin was five points or fewer and where the 2020 election was therefore also expected to be close (average absolute error of 4.5). In less competitive states the average absolute error was higher, and it was especially high in states that Trump won by more than five points in 2016. The average absolute error was 6.2 points in states Trump won by at least five points in 2016, which compares to 4.9 points in states Clinton won by more than five points. This difference in average absolute error between the polls done in these two groups of states was statistically distinguishable from zero at conventional levels (p=0.02).

According to the distribution of signed error (right-hand side of Figure 9), margins were too favorable to Biden regardless of the how competitive the state was in 2016. However, the largest overstatement of the Democratic-Republican margin occurred in states with a higher percentage of Trump voters. The average signed error of 5.3 points in states Trump won by more than five points in 2016 is considerably higher than the 3.9 points in competitive states and the 3.4 points in states Clinton won by more than five points. For polls in states won by Trump or Clinton in 2016 by more than five points, the difference in average signed error is statistically distinguishable from zero at conventional levels (p=0.03). The largest overstatement of the Democratic-Republican margin occurs in states Trump won handily in 2016, i.e., those states with a larger proportion of Trump supporters.

To describe the relationship between 2020 polling error and the level of Trump support in 2016, Figure 10 plots the average signed error in each state against the support Trump received in 2016. Although the polls overstated Trump support in some states (e.g., LA, MS) and were nearly exactly right in other states (e.g., GA, NM, CO, MD), the polls in many states significantly overstated the Biden-Trump margin. In states where Trump won 55% or more of the vote in 2016, the average signed error was often considerable. In the states of MT, UT, NE, TN, KY, IN, ID, ND, and WY, for example, the polls overstated the Democratic-Republican margin by 10 points or more.

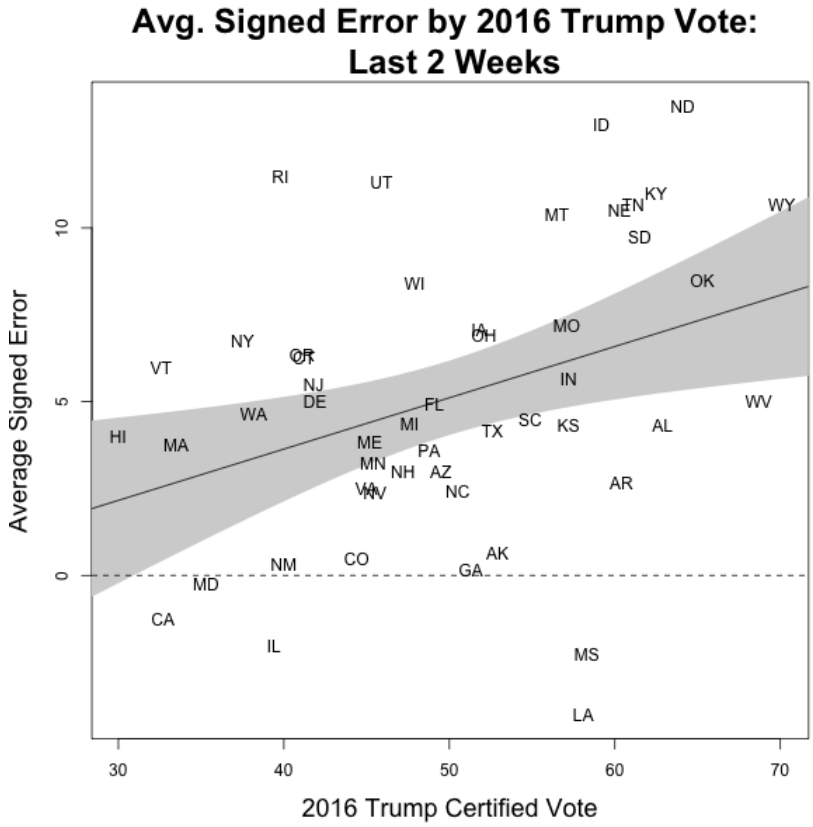


Figure 10. Average 2020 State-Level Signed Error by 2016 Trump Vote Share. The sloped line represents the simple regression line of average signed error on 2016 Trump vote. (Intercept = – 2.3 (std. err. 2.8), slope = 0.15 (std. err. 0.05) , and $R^2 = 0.13$)

Figure 10 confirms that the largest average polling errors in 2020 occurred in states with a larger percentage of 2016 Trump voters, but what this indicates about the source of the polling error is unclear. If Trump supporters were less likely to respond to polls than Biden supporters then larger errors would occur in places with more Trump support, exactly as observed. On the other hand, if Biden supporters were more likely to respond in those states, perhaps because they were enthusiastic about expressing their displeasure with Trump, then this too could have produced the same correlation, exactly as observed. This observation and association helps describe where larger polling error occurred but cannot fully explain the source of polling error.

7. Performance of 2020 Polls by State

Polls in most states overstated the Biden-Trump margin. Even within states, polls overstated the Democratic-Republican margin in senatorial and gubernatorial contests more than they overstated the Biden-Trump margin. The polling error was larger the more support Trump received in 2016 even after accounting for other between-state differences plausibly related to polling error.

The correlates of polling error by state can be examined by exploring whether or not the average polling error in a state correlates with a state's political environment.

Figure 11 begins this exploration by plotting the average signed error for each state. It is immediately obvious that polls overstated the Biden-Trump margin in nearly every state. Polls overstated the relative support for Trump in only a handful of states.

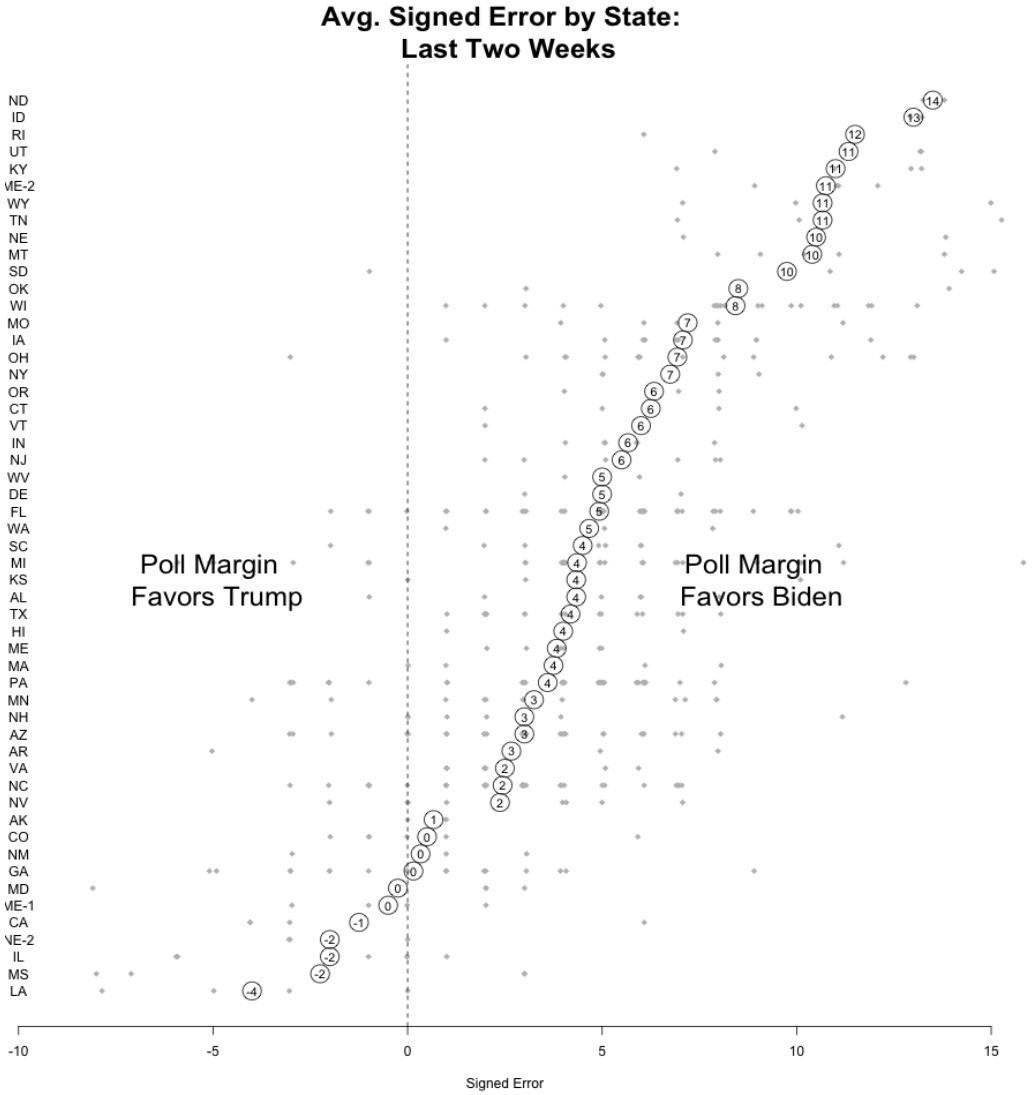


Figure 11. State-Level Average Signed Error by State. Circles denote the average signed error for each state. The signed error of individual polls in each state are plotted in grey points.

In nearly every state the average signed error is positive, because poll margins were too favorable towards Biden in nearly every state. Consequently, the pattern of average absolute error by state in Figure 12 is very similar to the pattern in Figure 11. Unlike in prior elections, the nearly pervasive pro-Democratic polling bias results in similarly sized signed error and absolute error across states. Although individual polls in some states were more accurate, the overall average of the state-level average absolute error is nearly six percentage points (5.9).[29]

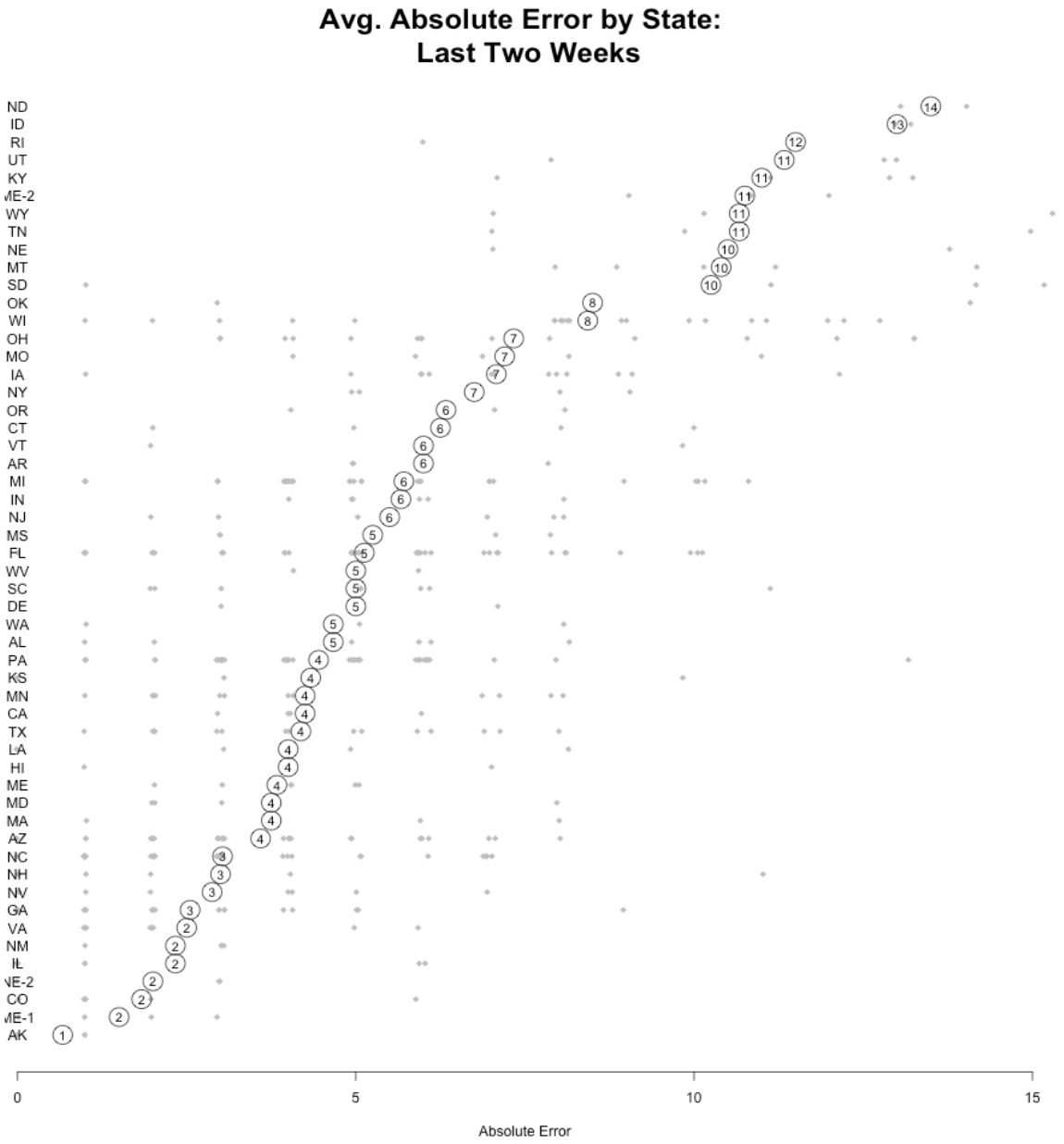


Figure 12. State-Level Average Absolute Error by State. Circles denote the average absolute error for each state. The absolute error of individual polls in each state are plotted in grey points.

[29] This is higher than the average absolute error of 5.1 points reported in Table 2 because 5.9 points is the average of the state averages. Put differently, the average absolute error of 5.1 is the average for all polls whereas the 5.9 points is the average of the state-level average of absolute error.

The overstatement of Democratic support was not unique to the presidential contest. Figure 13 graphs the average signed errors for senate contests (if any) as well as presidential contests for every state. Comparing the average state-level signed error among presidential state-level polls in states with (bottom) and without (top) a senate contest reveals two patterns.

First, the presence of a senatorial contest did not affect the polling error of state-level presidential polling. Polling error did not depend on whether other state-level races occurred.

Second, polls overestimated the Democratic-Republican margin in Senate races more than polls overestimated the Biden-Trump margin even within the same state.[30] More respondents chose a response other than the Democrat or Republican candidate in senatorial polls (an average of 6.1% of respondents) compared to presidential polls (4% of respondents), yet the difference in the average signed error seems hard to attribute to differences in respondent indecision, considering the magnitudes of each.

Overall, whatever caused the presidential polls to overstate the Biden-Trump margin in 2020 produced an even larger Democratic overstatement in senatorial polls.

[30] In other words, either the support for Democratic Senate candidates was more overstated relative to Biden, the support for Republican Senate candidates was more understated relative to Trump, or some combination of each.

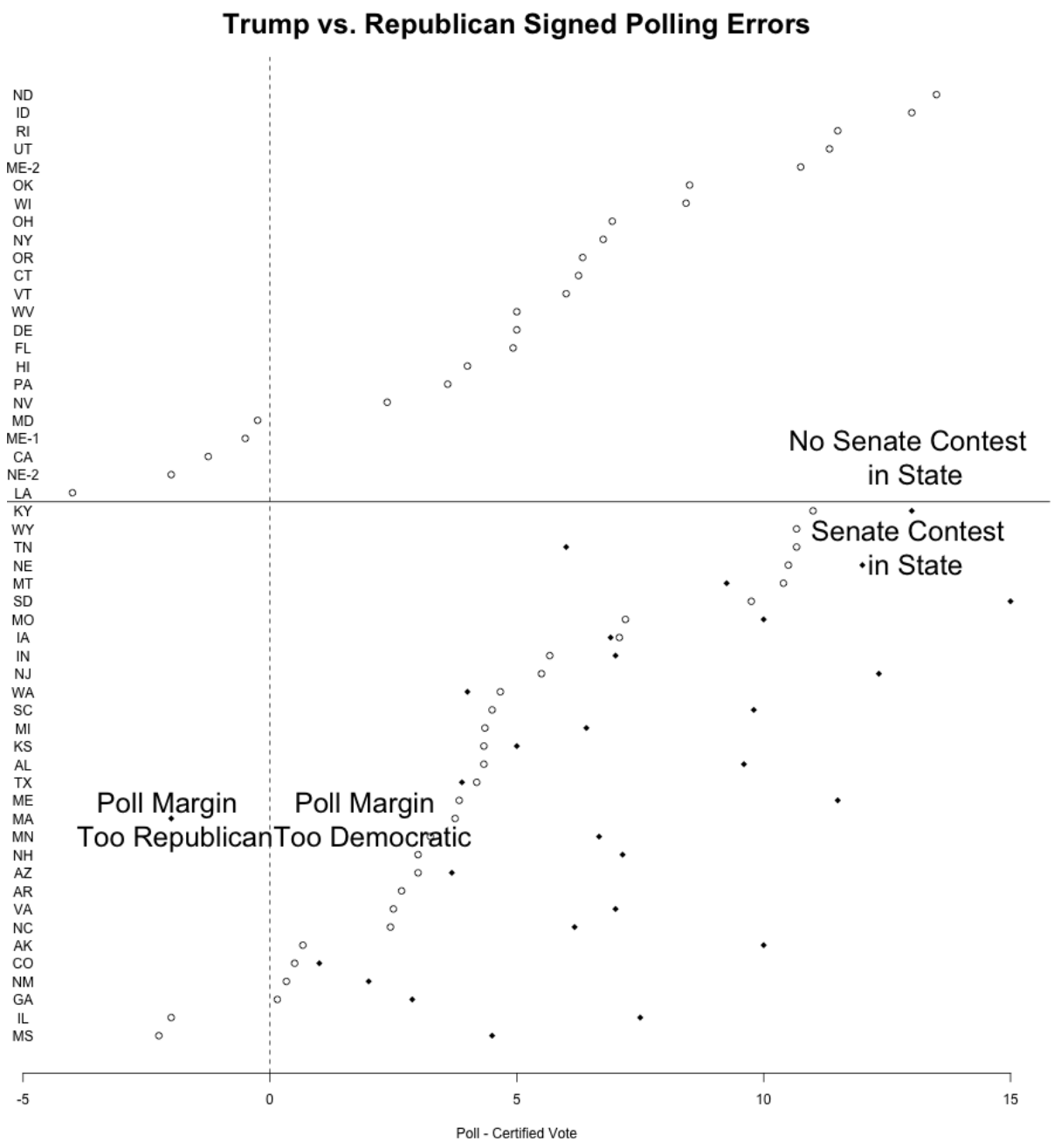


Figure 13. Average Signed Error for Presidential and Senate Polls by State. An open circle denotes the average signed error for state-level presidential polls, and a black solid dot denotes the average signed error for the senatorial polls in each state (if any).

7.1 Correlates of Average State-Level Error

Polls overstated the Biden-Trump margin more in states that Trump won by larger margins in 2016, even controlling for other between-state differences. All else equal, the overstatement of the Biden-Trump margin was also larger in states that were Whiter and less densely populated. States with a higher level of COVID-19 per capita had a slightly higher polling error as well.

The correlation between state-level polling error and features plausibly related to the difficulty of polling in the state were examined to further characterize the nature of state-level polling error in 2020. Examples include the racial diversity of the electorate in the state, the amount of early voting in a state, and the impact of the COVID-19 pandemic. The extent to which 2020 polling error is associated with 2016 support of Trump in the state, after controlling for other between-state differences, is of particular interest.

Many state-level factors that might have affected the 2020 polling error were examined, including the following: whether the state allowed same-day registration; early voting in 2020 compared to 2016; the ease of voting in the state using an index constructed to measure the costs of voting in the time of COVID-19 (Schraufnagel, Pomante, and Li 2020); the per capita incidence of COVID-19 as of Election Day; the percentage of Black residents; the percentage of residents with at least a bachelor's degree; the percentage of the state population over the age of 65; the log of the population density in the state; and the median income in the state.

Table 6 reports on this analysis of correlated variables. The results support the simplest explanation of the variation in polling error between states. (Table B2 in Appendix B shows similar results after controlling for the full set of covariates.) When predicting the *average signed error*, positive (negative) values indicate a larger overstatement (lower understatement) of the Biden-Trump margin; when predicting the *average absolute error*, positive (negative) values indicate correlates of larger (smaller) polling error.

	Descriptive Statistics	Regression Results	
	Mean (Std. Dev.)	Signed Error on the Margin (Standard Error)	Absolute Error on the Margin (Standard Error)
Trump's 2016 Certified Vote %	49.9 (10.1)	0.25* (0.07)	0.20* (0.06)
Percentage of the State that is Black	10.2 (8.8)	-0.33* (0.06)	-0.27* (0.06)
Log (State Population Density)	4.6 (1.4)	1.27* (0.43)	1.11* (0.38)
Percentage with a Bachelors' Degree	32.4 (5.2)	0.14 (0.12)	0.09 (0.10)
Total COVID-19 Cases Per Capita as of Election Day	0.03 (0.01)	102.6* (42.9)	115.9* (36.5)
Intercept		-17.0 (6.6)	-12.95 (5.9)
Number of polls		49	49
Adj. R^2		0.48	0.53

Table 6. Regression of Signed Error and Average Absolute Error on Characteristics of States and State Presidential Polls Conducted in the Final Two Weeks. An asterisk (*)denotes statistical significance at 0.05 or better. The first column provides the mean and standard deviation for the state-level characteristics.

The results for the state-level average signed error reveals that one of the strongest predictors of state-level polling error is the 2016 state-level support for Trump. Even after controlling for other factors, a one-percentage-point increase in Trump's 2016 vote share is associated with a 0.25-percentage-point-larger signed error on the margin. All else equal, one standard deviation (10.1 percentage points) in the support for Trump at the state level correlates with a 2.5 percentage points (0.25 x 10.1) deviation from the average signed polling error. Put differently, the average signed error on the margin for polls is predicted to be 2.5 percentage points larger in states where Trump received 55% of the vote in 2016 compared to the average signed error in states where Trump received 45% of the vote in 2016.[31]

[31] The importance of Trump's 2016 state-level certified vote percentage in predicting the variation of polling error across states can be appreciated by re-estimating the models in Table 6 without that measure. Recalculation results in adjusted-R^2 values of 0.33, 0.28, and 0.31 for regressions predicting the average signed error on the margin, average Trump error, and average Biden error, respectively. Some of the between-state variation in polling error can be predicted using other factors but better predictions are obtained by accounting for Trump's 2016 vote share.

Controlling for Trump support, polling error was larger in Whiter and more rural states. A one-standard-deviation increase in the percentage of the state population that was Black (i.e., an 8.8-percentage-point increase in the size of the Black population) was associated with a 2.9-percentage-point decrease in the signed error and a one-standard-deviation increase in log population density increases the average signed error by 1.8 percentage points. The signed error was also higher in states that were more affected by the COVID-19 pandemic as of Election Day. A one-standard-deviation increase in cases per capita is associated with a 1.1-percentage-point increase in signed error.[32]

Predicting the absolute error produces similar findings. Considering that the signed errors favored Biden in nearly all states and that there were so few state errors in the other direction, the similarities are not surprising.

That the polls overstated Biden's support more in Whiter, more rural, and less densely populated states is suggestive (but not conclusive) that the polling error resulted from too few Trump supporters responding to polls. Similar to patterns described earlier, a larger polling error was found in states with more Trump supporters. These correlations are consistent with a polling error caused by too few Republican-supporting respondents (or by Republican respondents indicating support for Democrats and then voting for Republicans) but these correlations cannot conclusively identify the underlying cause. For example, polling error might have been caused by unaffiliated voters (including unaffiliated new voters) motivated to participate in 2020 because of Trump. Those voters may have may have been less likely to respond to polls, following Trump's exhortations to ignore "fake polls." The observed patterns could be produced if those voters were more likely to live in states that were more rural, Whiter, and supported Trump in 2016; the same pattern would result even if the percentages of Democrats, Republicans, and independents polled correctly matched the 2020 electorate.[33]

[32] Strengthening the inference of a real COVID-19 pandemic effect requires comparing the survey error within a state over time to better leverage the available variation.

[33] Alternatively, even if the Republicans and Independents who responded were representative of those who did not (and Republican supporters participated at the same level as they have in the past), if Democrats were motivated to participate more than usual and their participation was highest in places were Trump did the best in 2016, this could also produce the patterns we observe.

8. Performance of Polls: Candidate Error

The average topline support for Biden and Trump in polls overstated the support for Biden relative to his actual vote share and understated the support for Trump. This characterization depends on whether we include or exclude the 4% of respondents who chose a response other than Biden or Trump. How pollsters report their results can affect impressions of the polls.

A Democratic-Republican margin that is too favorable to Biden can be the result of overestimating support for Biden, underestimating support for Trump, or some of each given the symmetry in a two-candidate race. To characterize polling error we investigated the average candidate error in state-level presidential polls.

Candidate error is measured using the candidate's percentage in the poll minus the candidate's percentage in the certified vote. For example, if Trump received 54% in a poll and 58% in the certified vote, the candidate error for the Trump in the poll would be – 4. Even though only a small percentage of respondents (on average 4% in state-level presidential polls conducted in the last two weeks) provided a response other than Biden or Trump (e.g., "don't know"), whether the respondents who choose an answer other than Biden or Trump are included or not affects the magnitude of the candidate error.

We focused on the topline poll results for Biden and Trump rather than the two-candidate support because topline percentages typically are the ones discussed in poll coverage. Such coverage may have led people to interpret the topline polling percentages as predictions of the certified vote share. Insofar as this is true, comparing the topline poll results with the certified vote share highlights how poll results often are reported and interpreted in ways that shape public perceptions and expectations.

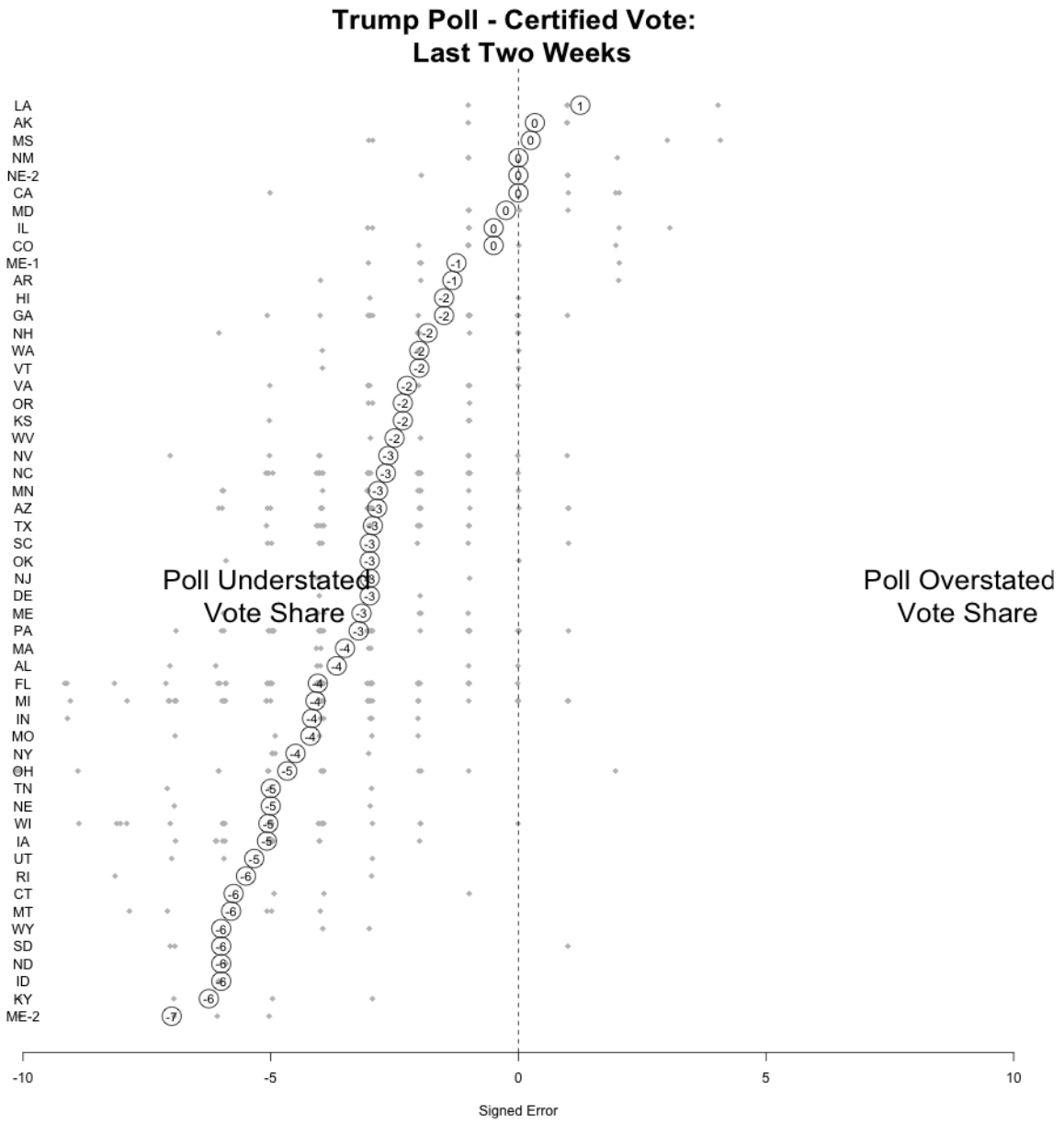


Figure 14. Polling Error on Trump by State. All polls with a field period ending between 10/21/20 and 11/3/20 are included. Circles denote the average polling error on Trump's vote share for each state (as reported within each circle). Trump error is computed as Trump's percentage in the poll minus Trump's percentage in the certified vote. Results are sorted by the average Trump error in the state. Trump error for individual polls are plotted in grey points.

Figure 14 shows that the average support for Trump in the polls understated Trump's certified vote share in nearly every state. Overall, the average difference between Trump's support in the polls and his certified vote was – 3.3 percentage points, indicating that the support Trump received in the topline poll results was 3.3 percentage points lower than the vote share Trump received in the certified vote. Moreover, for polls that reported a margin of error, the difference between the reported percentage of voters supporting Trump and Trump's certified vote share in the state was greater than the margin of error in 44% of the state-level presidential polls. The fact that the error on the proportion was often larger than the margin of error again highlights the inadequacy of the margin of error as a quantification of uncertainty and error.

The polls did a better job estimating the support for Biden (Figure 15). In several states, e.g., TN, RI, UT, ID, and ND, Biden's estimated support was much higher than the certified vote; but the difference between Biden's average support in the polls and the certified vote was often within a percentage point or two. Across all polls, the average candidate error for Biden was 1.0, indicating that the topline polling number for Biden was one percentage point higher than what Biden received in the certified vote. Moreover, for 86% of the polls that reported a margin of error, the difference between the support for Biden in the polls and his certified vote share was within the margin of error.

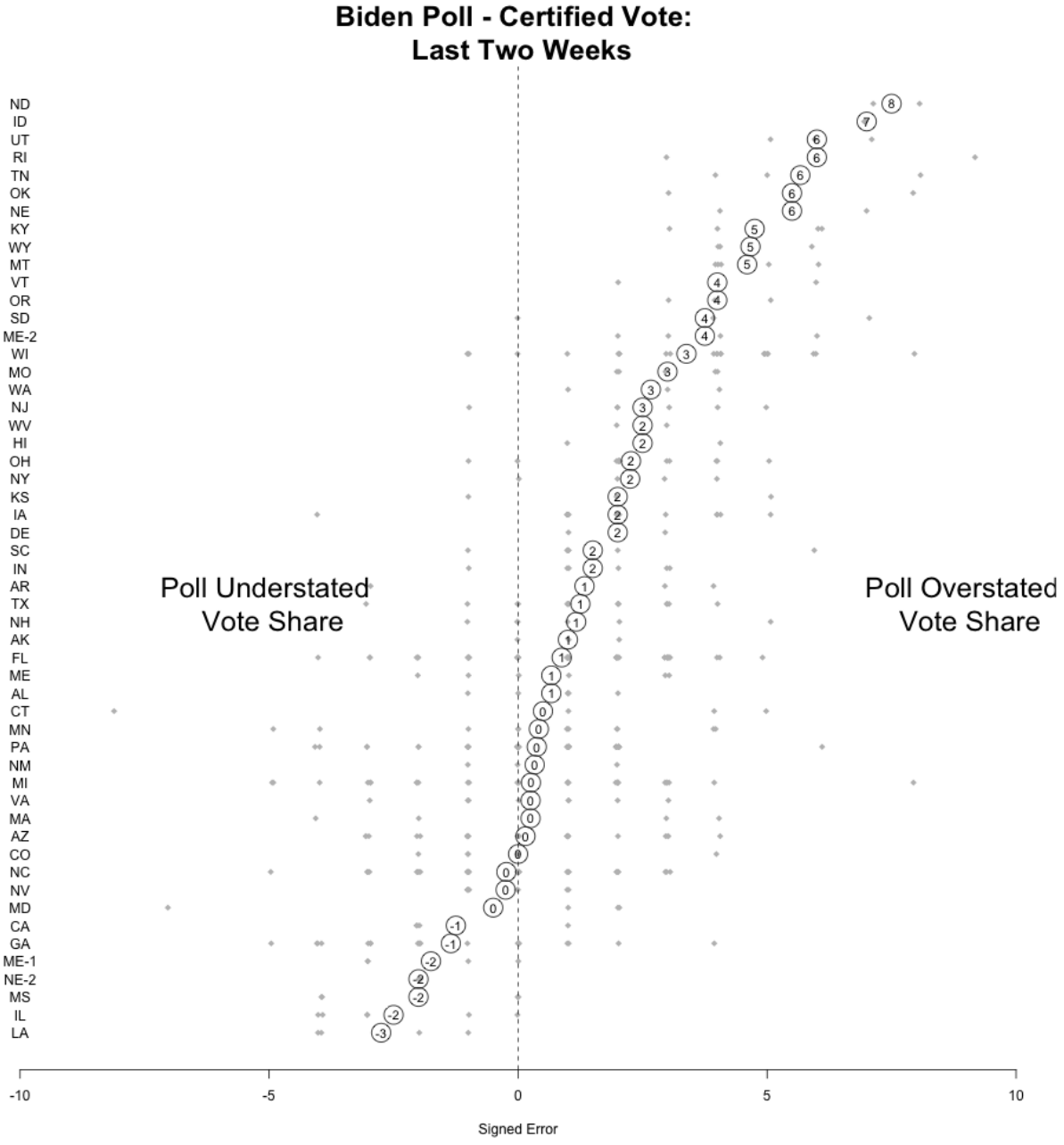


Figure 15. Polling Error on Biden by State. All polls with a field period ending between 10/21/20 and 11/3/20 are included. Circles denote the average polling error on Biden's vote share for each state (reported within each circle). Biden error is computed as Biden's percentage in the poll minus Biden's percentage in the certified vote. Results are sorted by the average Biden error in the state. Biden error for individual polls are plotted in grey points.

The patterns in Figures 14 and 15 are consistent with the claim that the polls had too few Trump supporters, but they cannot prove that this was the case and it is important to not overinterpret the patterns. Depending on how we treat the 4% of respondents who give a poll response other than "Biden" or "Trump," the characterization of the candidate error changes.

As Figures B2-1 and B2-2 in Appendix B reveal, removing the 4% of respondents who gave an answer other than Biden or Trump to focus on the two-candidate vote increases the support for each candidate in polls by changing the denominator. Comparing the two-candidate support in each poll to the certified vote reveals that 86% of polls had Trump's support within the margin of error, and 65% of polls had Biden's support within the margin of error. Looking at the two-candidate support relative to the two-candidate vote share (that is, "percentaging out the undecideds") leads to a conclusion that the polls understated Trump's certified vote share by 1.4 points and overstated Biden's certified vote share by 3.1 points. In other words, using the two-candidate support rather than the topline polling result would change our conclusion about whether the larger contribution to polling error is the overstatement of Biden support or the understatement of Trump support.

Even though it is impossible to resolve whether the larger source of polling error is an overstatement of Biden support or an understatement of Trump support based on this analysis, the differences in the conclusions that are reached highlight how the decision to report a topline number based on all respondents or based on the two-candidate share can affect how the performance of the polls is perceived and evaluated. Poll coverage of topline polling results likely creates an expectation that the reported toplines can be used to estimate vote shares and diagnose polling error even though the topline support is affected by the number of response options that are given to respondents (e.g., the presence of "Don't know" or "Other").

Our characterization of how well the polls performed for each candidate depends on whether we use the topline support or the two-candidate support. This fact highlights the importance of seemingly small decisions on how the results of polls are interpreted and characterized (Mitofsky 1998). It also highlights the difficulty of using poll results to diagnose what actually produced the polling error in 2020, a task to which we now turn.

PART TWO: EXPLAINING WHAT HAPPENED

Having described the overall performance of the polls and how it varied or not across different types of polls and contests, we now examine the factors that can and cannot explain the patterns documented in part one.

Explanations unfortunately can be complicated and definite conclusions are often impossible. Certain data that are essential to fully assess what happened can be unknown (e.g., how pollsters weighted their data to account for nonresponse, identify likely voters, etc.) or unknowable (e.g., what the electorate actually looked like and how poll respondents compared to nonresponders).

Without knowing the number, opinions, and vote choices of voters not included in polls, either because they were not included in sampling frames or because they chose not to participate, it is not possible to conclusively identify the source(s) of polling error. Given these limitations, the initial objective is to eliminate several potential explanations of polling error. By narrowing the set of explanations, the remaining explanations for polling error can be identified as well as the patterns that the errors must explain.

The patterns in part one point to some causes of polling error that can be immediately eliminated as follows.

- *The underlying understatement of Republican support or overstatement of Democratic support almost certainly did not result from questions about a specific candidate.* The polls favored Democrats at the expense of Republicans regardless of interview mode and regardless of contest. Respondents lying to pollsters about their support for Trump, for example, would not explain why the error was larger for Republican candidates for governor and Senate.

- *The polling error was not caused by errors in the sampling frame due to the exclusion of new voters.* Polling error was observed regardless of how respondents were sampled or contacted; that is, it was similar for online, RBS, and RDD surveys.

- *Polling error was not due to last-minute changes or shifts in voter sentiment.* The size and nature of polling error among polls conducted in the last two weeks was similar to the polling error among polls conducted in the last three days. Whatever the cause of the larger-than-usual polling error may be, it affected polling throughout the election.

In the next five sections, several additional possible explanations are considered and eliminated or not as follows.

- Section 9 shows that polling error in 2020 was not due to certain issues that affected polls in 2016. There was no evidence of late deciders causing polling error, and most 2020 polls accounted for the education of respondents when weighting for nonresponse.

- Section 10 shows that polling error in 2020 was not caused by Biden supporters reporting that they had voted when they had not. The expanded use of in-person early voting and mail voting (i.e., early/mail voting) did not overrepresent Biden supporters by reason of respondents who falsely claimed they voted.

- Section 11 investigates how alternative voting methods might underrepresent Election Day voters. Based on the data available to us, this did not appear to be the case.

- Section 12 considers the effects of the historic turnout and the extent that new voters posed challenges for polling. New voters cannot be eliminated as a source of polling error.

- Section 13 builds on the analysis of Section 12 by considering a bundle of potential issues, including coverage/sampling (e.g., were Republican supporters less likely to be contacted and/or Democrats more likely to be contacted?); between-party nonresponse (e.g., were Republican supporters were less likely to respond than Democratic supporters?); and within-party nonresponse (e.g., were the Republicans who responded more likely to support Democrats than Republicans who did not respond?). The analysis points to several patterns, but the conclusions depend on many assumptions. It is impossible to identify who did not respond and why by looking only at who did respond; information on those who voted but did not respond either is not yet available or does not exist.

9. What About the Issues Identified in the 2016 Report?

The 2016 Task Force report identified three principal sources of polling error.[34] First, late-deciding voters and earlier-deciding voters behaved differently, which may have caused an understatement of Trump's support in pre-election polls. Second, weighting to account for differences by education level may have understated Trump's support, especially among White voters. Third, assumptions relating to turnout and the composition of the electorate turned out to be incorrect, which is discussed in later sections.

This section addresses whether or not the first two issues identified in the 2016 Task Force report contributed to polling error in 2020.

9.1 Error Due to Late Deciders: Not an Issue in 2020

> *The number of late deciders is not large enough and their opinions are not distinctive enough to cause the overstatement of the Biden-Trump margin.*

In 2016, late-deciding voters supported Trump at a much higher rate than early-deciding voters, especially in critical Midwestern states. A sizable fraction of the 2016 electorate decided in the final few days, according to exit polls, and there were enough late-deciding Trump supporters to account for some of the observed polling error.

This was not an issue in 2020. Early in-person voting and voting by mail were prevalent in 2020 due to the COVID-19 pandemic and, as a result, many voters decided well before Election Day. As indicated in Table 2, almost every respondent answered either "Biden" or "Trump" when polled. Even if all of the 3.5% of respondents who answered something else ultimately voted for Trump, that would not account for the observed polling error in 2020.

[34] https://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx

Exploring further and following the lead of the 2016 Task Force, the size and vote choice of late-deciding voters was examined using the National Election Pool national exit poll. Fewer than 5% of respondents at exit polls said they decided in the final week, consistent with Table 2.

Even though voters who decided "in the last week" were more likely to support Trump over Biden (by 16 points), they shifted the margin by less than a percentage point since they made up only 5% of the electorate (.05 x .16 = .008). There is no evidence that late-deciding voters caused error in estimates from the national polls in 2020. Table 7 shows that late-deciding voters in each state not only make up a comparatively small fraction of the overall electorate, but also were more likely to report voting for Biden over Trump. Only in Wisconsin did the voters slightly break for Trump in the final week, and the size of the effect in Wisconsin was less than two-tenths of a percentage point (.05 x .03 = .0015).

State	Avg. Signed Polling Error	% of Voters Deciding in Final Week	Margin Among Voters Deciding in Last Week	Margin Among Voters Deciding Earlier Than the Last Week	Difference
National	+4	5	Trump +12	Biden +4	Trump +16
Florida	+5	4	Biden +4	Trump +1	Biden +5
Georgia	0	4	Biden +16	Biden +1	Biden +15
Michigan	+4	4	Biden +16	Biden +5	Biden +11
Minnesota	+3	6	Biden +18	Biden +10	Biden +8
North Carolina	+2	3	Biden +8	Trump +1	Biden +9
Texas	+4	3	Biden +9	Biden +3	Biden +6
Wisconsin	+8	5	Trump +3	Tie	Trump +3

Table 7. Vote of Late Deciders in the Last Week in Critical States. Results based on the final weighting of the National Election Pool's exit poll. Average signed error is from all public polls conducted in the last two weeks.

According to the exit polls, few voters decided in the last week and many supported Biden over Trump; therefore it is unlikely that the polling error in 2020 was caused by late-deciding voters.

9.2 Error Due to a Failure to Account for Educational Differences: Not an Issue in 2020

Unlike 2016, almost all polls in 2020 accounted for education level when weighting their results. Our analysis in Section 12 below reveals no evidence that the weighting targets were incorrect.

Another issue identified by the 2016 Task Force report was that only a handful of state-level polls accounted for education-related differences when weighting their results to account for nonresponse and coverage issues. Voting choices in 2016, particularly among White voters, were highly correlated with educational attainment. College graduates heavily favored Clinton in 2016 and non-college graduates heavily favored Trump. The failure to account for education differences was a problem for polling in 2016 because individuals with more education were also more likely to respond to polls relative to those with less education, resulting in an overrepresentation of Clinton supporters among the weighted respondents.

Almost all state-level, pre-election polls followed the advice of the 2016 Task Force report and weighted the results using education. Among the 438 state-level polls conducted within the last two weeks of the election, weighting parameters could be identified in 317 polls; among those 317 polls, 292 (92%) used education in their weighting methodology, and only 25 (8%) did not.[35] Whether or not the weighting targets used to weight education level were correct is impossible to know, but most polls attempted to account for nonresponse error correlated with educational attainment by weighting responses using education.

Nearly every state-level presidential poll in 2020 weighted responses by the education level. Since the recommendations of the 2016 Task Force report were adopted by the industry, the issues affecting polling in 2020 went beyond what accounting for education alone could correct. Although incorrect assumptions about education levels could possibly account for part of the 2020 polling error, the analysis presented in Section 12 does not identify this factor as a primary source of polling error.

10. Error Due to Differences in Self-Reported Intent to Vote: Unlikely to Be a Factor

Biden supporters and Trump supporters were equally likely to vote after reporting that they would. Overstatement of the Biden-Trump margin was not a result of Biden supporters saying that they would vote and then failing to do so.

Another potential explanation is based on unequal numbers of Biden and Trump supporters who did not vote after saying they would. Democratic supporters being more likely than Republican supporters to overreport their likelihood of voting could explain the overstatement of the Democratic-Republican margin in the polls. This error would include not only self-reported intention to vote, but also the self-reported behavior of respondents who claimed they already voted.

[35] Some polls that did not to use education as a weighting factor instead used geographic divisions within the state (or urban/rural classifications) or income.

The accuracy of self-reported voting intention and behavior was especially relevant in 2020 because of the politicization of voting methods. Pre-election polls typically only count those who say they are extremely likely to vote and those who say they already voted, or they rely on past voting behavior. Democrats were more likely than Republicans to vote by mail or vote in person early, and Republicans were more likely than Democrats to vote on Election Day. These differences could affect the accuracy of polls, according to whose responses were counted by polls. For example, respondents who said they had already voted in-person early or by mail might always be counted in the polls, while Election Day voters might only sometimes be counted according to their likelihood of voting. Mistaken self-reports of Biden supporters could result in too many Biden supporters being included in the polls.

For the 2020 election, there is no evidence that respondents who supported Biden were more or less likely to vote than respondents who supported Trump. On the contrary, an AAPOR Transparency Initiative member organization measured the validity of the self-reported vote of RBS poll responses. Results shared with the Task Force revealed that Biden and Trump supporters were equally likely to vote. If they are equally likely to vote when they say they will vote, then it seems unlikely that any combination of likely-voter models and self-reporting error could produce the polling error in 2020.

Table 8 reports on polls matched to a voter file using the voter-file crosswalk (*Pct. Matched*). This set of well-matched polls includes states with large (e.g., MI, PA, and WI) and small (e.g., GA, NC) polling error. Of primary interest is whether the accuracy of the self-reported voting intention of Biden supporters is lower than the accuracy of the self-reported voting intention of Trump supporters.

	Sample Size (Report Intention to Vote)	% Sample Matched with Records	% Intending to Vote Who Did	% of Biden Supporters Who Voted	% of Trump Supporters Who Voted
Arizona	1,252	99	92	94	93
Florida	1,451	99	92	94	94
Georgia	759	84	99	88	82
Iowa	753	99	90	92	92
Kansas	755	98	88	93	89
Michigan 1	856	98	82	83	85
Michigan 2	1,231	99	89	90	89
Minnesota	1,005	98	93	93	93
Montana	758	99	95	98	94
Nevada	809	99	87	92	89
North Carolina	758	99	95	98	94
Pennsylvania	1,862	99	91	93	92
Wisconsin	1,253	98	95	96	95

Table 8. Accuracy of Self-Reported Vote Intention. Results are from comparing self-reported vote intention or report of early voting to voter file. Michigan 2 and Minnesota use a different pollster and voter file. The percentages are unweighted but calculations using the weighted proportions reveal similar patterns.

There is no evidence that Biden supporters were any less likely to vote than Trump supporters. In Arizona, for example, 94% of Biden supporters included as a likely voters did vote, and 93% of Trump supporters who reported they were going to vote actually voted. If anything, self-reported Biden supporters were slightly more likely to vote than Trump supporters.

The lack of differences in Table 8 could mask variations in the accuracy of self-reported intention to vote, depending on the self-reported method of voting. This possibility was investigated by cross-checking the voting method against the percentage of respondents who were confirmed to have voted for each candidate in the states of Pennsylvania, Michigan, and Wisconsin.

Table 9 reports on the percentage of respondents validated as having voted (*% Voted*) for each voting method. It also states how large that group was relative to the set of weighted respondents who indicated that they supported Biden or Trump.[36] Although some differences emerge, polling error that is due to absentee and in-person early voters exaggerating their likelihood of voting for Biden is not supported by evidence.

	% Validated as Having Voted	% of Weighted State Sample
Michigan		
Election Day Voters: Trump Supporters	83	30
Election Day Voters: Biden Supporters	69	11
Absentee Voters: Trump Supporters	86	9
Absentee Voters: Biden Supporters	88	31
In-Person Early Voters: Trump Supporters	84	7
In-Person Early Voters: Biden Supporters	90	12
Pennsylvania		
Election Day Voters: Trump Supporters	93	40
Election Day Voters: Biden Supporters	94	20
Absentee Voters: Trump Supporters	93	8
Absentee Voters: Biden Supporters	87	33

[36] Thus, only responses of "Biden" or "Trump" are counted. These are compared with the weighted percentage of Biden or Trump supporters who indicated that they were voting by each method relative to the total (weighted) number of Biden and Trump supporters in the state. This approach removes undecided and third-party supporters from the denominator and weighting and shows the robustness of the earlier conclusions.

Wisconsin		
Election Day Voters: Trump Supporters	94	28
Election Day Voters: Biden Supporters	93	11
Absentee Voters: Trump Supporters	95	24
Absentee Voter: Biden Supporters	96	40
In-Person Early Voters: Trump Supporters	95	10
In-Person Early Voters: Biden Supporters	94	17

Table 9. The Percentage of Biden and Trump Supporters Who Voted by a Self-Reported Voting Method in Three States. The percentages incorporate respondent weights. The size of each group is relative to the weighted number of respondents who choose either Biden or Trump.

Table 9 reveals only small differences between Trump and Biden supporters. Most differences are within a percentage point of one another.

The largest discrepancy in Table 9 is among Biden Election Day voters in Michigan. Although only 69% were validated as having voted, the effect is small for several reasons: only 11% of the sample were Election Day Biden supporters and Biden supporters were 14 percentage points less likely to actually vote than Trump supporters; hence the overall margin is shifted by only 1.5 points (.14 x .11 = .015) relative to the validated vote margin. This small amount can matter in close elections, but there does not seem to be a systemic pattern of differential self-reporting error that would explain the tendency of the 2020 polls to overstate the Democratic-Republican margin.

An important caveat is that only one set of polls was examined. The types of likely voters in these polls could differ from the types in other polls. That no differences were found in the voting behavior of respondents in these polls does not preclude the possibility that these factors would account for an importance source of error in different polls that use other methods of selecting respondents and identifying likely voters.

11. Error Due to Method of Voting and Likely Voters: Unlikely to Be a Factor

> *The relative percentages of Election Day and pre-Election Day voters were quite accurate for the polls conducted close to Election Day. The polling error was not obviously due to too few (weighted) Election Day voters. The overstatement of the Biden-Trump margin was largest among those who reported that they were voting prior to Election Day. For polls conducted more than two weeks prior to Election Day, the self-reported method of voting was not very accurate.*

Pre-election polls in 2020 faced more uncertainty than usual in estimating the likely electorate because many states changed how they allowed their citizens to vote due to the COVID-19 pandemic. Party leaders also politicized the various methods of voting: Democratic voters were encouraged to vote early via mail or in person and Republican voters were urged to vote in person before or on Election Day. If polls predicted too many early voters and too few Election Day voters, then the overstatement of the Democratic-Republican margin could have been replicated.

To determine if the polls contained the correct proportion of voters by method of voting, the percentage of votes cast using each method of voting, according to certified vote reports and voter files, was compared to the proportion of self-reported voting intention among the polls. Unfortunately, not every poll reported the required information, and there were often notable differences in how polls asked about the method of voting.[37]

Among the 23 state-level presidential polls conducted in the final two weeks, the signed error on the percentage of respondents who reported using the different methods of voting, on average, was accurate. (See Table 10.) In other words, averaging across polls and states to provide enough observations for analysis reveals that the proportion of voters who indicated that they were voting on Election Day or before Election Day closely matched the proportion of votes that were cast using those methods.

Even when average absolute error is used to avoid the cancellation of offsetting errors (i.e., if some polls overestimated the size of the Election Day vote and others underestimated the size of the Election Day vote, the average of those errors may be zero because the errors cancelled each other out) the average absolute error is modest. Among polls conducted during the last two weeks, the average absolute error on the size of Election Day voters was four percentage points and the average absolute error on the size of early/mail voters was five percentage points. Given the small size of these errors, it can be concluded that neither the overstatement of Democratic support nor the understatement of Republican support in the polls was caused by polls predicting too few Election Day voters or too many mail-in or in-person early voters.

	Signed Error		Absolute Error	
Voting Mode	Size	Margin	Size	Margin
Last Two Weeks (N = 23)				
Early/Mail	0	+8	+5	+9
Election Day	0	-3	+4	+8
All Polls (N = 42)				
Early/Mail	-7	+15	+10	+16
Election Day	+6	-4	+9	+9

Table 10. Error in the Size and Vote Choice of Early/Mail and Election Day Voters Pooled Across States. This table pools polls in Georgia, North Carolina, Pennsylvania, Iowa, South Carolina, and Texas because those were the only states with multiple polls reporting breakouts in the last two weeks. A positive signed error indicates an overstatement by the polls.

[37] The polls containing the individual-level data analyzed in Section 9 are included in this analysis.

Disaggregating the signed error and absolute error in these polls to calculate the polling error by method of voting reveals the polls overstated the Biden-Trump margin by eight points among self-reported early/mail voters and understated the Biden-Trump margin by three points among self-reported Election Day voters. In terms of the average absolute error, the error was nine points among self-reported early/mail voters and eight points among self-reported Election Day voters. While strong conclusions are impossible given the limited data, the overstatement of Biden-Trump support seems to be concentrated among respondents who reported that they would vote prior to Election Day.

Although not obviously relevant for polling error in 2020, note the sizes of the errors for polls conducted more than two weeks prior to Election Day. (See the bottom two rows of Table 10.) When all 42 of the polls are included in the analysis, respondents self-report that they are more likely to vote on Election Day (by six percentage points) and less likely to report voting early or by mail (by seven percentage points) compared to the proportion of the electorate that actually did. Whether these discrepancies were because of the new voting procedures used during the pandemic or because respondents were unsure about how they would vote, it is worth noting that early poll results provided a misleading portrait of how respondents intended to vote. Going forward, it seems wise to treat early self-reported methods of voting with caution.

12. Error Related to "New" Voters: Yes, in Part

> *The percentage of voters who voted in 2020 but not 2016 was large enough to produce much of the polling error. Biden support was greater among new voters than suggested by the difference between votes cast in 2020 versus 2016. Without knowing the size of this group of voters and how they voted, the difficulty of polling new voters cannot be stated conclusively to be the cause of the overall polling error in 2020, but available data suggest it cannot be ruled out as one possible cause.*

One explanation of polling error is the dramatic increase in the number of 2020 voters. Nationwide, 22 million more certified votes were cast in 2020 than 2016. New voters include newly eligible voters as well as newly energized voters who were inspired to participate by the candidacies of Trump or Biden. To the extent that the 2020 outcome was shaped by the presence of "new" voters who had not voted in 2016, those new voters could affect the accuracy of pre-election polls in 2020. [38]

A lower bound on the percentage of voters who voted in 2020 but not 2016 is obtained by assuming that every 2016 voter also voted in 2020. In that case, 14% of the certified votes cast in 2020 would be due to voters who did not vote in 2016 (22 million over 158 million). The actual percentage of voters who voted in 2020 but not 2016 is certainly higher since not every 2016 voter also voted in 2020. The additional voters in 2020 relative to 2016 could have had a large effect on the overall polling error.

We focus on two questions: Did the polls have the right proportion of new voters? And did the candidate preferences of the new voters who were included in polls reflect the preferences of those voters not included in the polls?

[38] Those who voted in 2020 but not 2016 are the main focus for the sake of analytical traction, recognizing also some who voted in 2012 and then sat out 2016 before voting in 2020.

It is a logical truth that the opinions of respondents cannot be compared with the opinions of nonrespondents; likewise, differences between new voters who did and did not respond to the polls cannot be determined. Nonetheless, several benchmarks were used to explore the effects of new voters on the 2020 polling error.

Several measures are used to estimate how many voted in 2020 but not 2016. As a lower bound, the difference between votes cast in 2020 versus 2016 can be as a percentage of votes cast in 2020 (*Lower Bound).* The assumption that every 2016 voter also voted in 2020 is clearly wrong, but this estimate puts a lower bound on the number of new voters. In Arizona, for example, 3.4 million votes were cast in 2020 and 2.6 million voters were cast in 2016. If every 2016 voter also voted in 2020, then 800,000 voted in 2020 but not 2016.[39] As a fraction of the total votes cast in 2020, roughly 23% of the Arizona electorate voted in 2020 but not 2016.

Besides this lower bound calculation, voter files were used to identify the number of voters who voted in 2020 but not 2016. Tracking registered voters over time can be difficult since they move and change their names. Two different proprietary data files were used to perform the calculation, given the difficulty of tracking voters (see, for example, Ansolabehere and Hersh 2014).

Table 11 lists various estimates of how many voted in 2020 but not 2016. In nearly every state, the voter-file-based estimates greatly exceeded the lower bound, vote-based calculations. Highlighting the difficulty of the task, even many months after the election and final certification, it remains unclear how many “new voters” voted in 2020.

Critical for our purposes, the fraction of new voters is considerable. It is often large enough to affect the overall polling error in 2020. If the Voter File 1 estimate is accurate, for example, and 34% of Michigan voters voted in 2020 but not 2016, then size of this group in the overall electorate means that errors made when polling this group could greatly affect the overall polling error. If the new voters were three points more likely to support Biden and three points less likely to support Trump, then a six-point error on the candidate margin among a third of the electorate would shift the overall error on the margin by two points.

It was even more difficult to determine if the new voters included in polls reflected new voters who were not. To provide a benchmark for how the new voters may have voted, 2020 voters must be compared to 2016 voters. In particular, support among new voters can be estimated by comparing the change in Biden/Clinton votes (from 2016 to 2020) to the change in Trump votes in the same period. In Arizona, for example, comparing 2020 to 2016, Biden received 511,000 more votes and Trump received 409,000 more votes. The difference of these differences (102,000) can be expressed as a fraction of increase in votes going to either Trump or Biden (920,000). This calculation suggests that Biden’s margin must have been 11 percentage points larger among these voters.[40] If every voter who voted in 2016 voted the same in 2020 as 2016, how must the new voters vote to produce the difference in results between 2020 and 2016? Such calculations are crude benchmarks at best, yet they arguably help to ballpark expectations. The results of these calculations are reported in Table 11 as the *Vote Margin (Est)*.[41]

[39] Of course voters can move between states, i.e., some "new” voters may have voted in 2016 in a different state.
[40] The denominator here is the sum of the votes for Trump and Biden (920,000). There are 920,000 “new” voters but only 780,000 new votes when accounting for third-party candidates in 2016.
[41] While Biden's margin among 2016 nonvoters cannot be known, these figures serve as a rough benchmark under the assumptions that neither Trump nor Biden disproportionately *netted* votes from vote switching or differential turnout. To be clear, our approximation of Biden's margin based on increases in certified totals isn't technically correct because it could also result from vote switching and differential turnout of Clinton and Trump 2016 voters. While approximate, it is hard to imagine that these factors could explain the 20-point differences between this calculation and the poll results.

In terms of size and support, benchmarks could also be compared to the results of polls that asked whether respondents had voted in 2016. Table 11 reveals that the percentage of 2020 poll respondents who said they did not vote in 2016 (*% Not Voting in 2016*) was often closer to the back-of-the-envelope calculations (*Lower Bound*) than to the voter-file estimates (*Voter File 1 Size*, *Voter File 2 Size*) regardless of the voter file used. In Michigan, for example, the percentage of poll respondents who indicated that they did not vote in 2016 was 14%, which is close to the 13% estimated using the difference in certified votes being cast but less than half the size of the voter-file estimates.

It was not possible to ascertain whether or not RBS polls reflect larger trends, but these results certainly suggest that the fraction of voters who voted in 2020 but not 2016 was too small in the polls that were analyzed. In Wisconsin, for example, the percentage of voters who voted in 2020 but not 2016 was less than half of what the two voter files suggest. Among these polls, the proportion of new voters appeared to be considerably understated.

A similar disconnect is revealed in determining if vote choices of new voters closely match best estimates the same. Respondents who indicated that they were voting in 2020 and had not voted in 2016 were far more likely to support Biden than Trump. This was true regardless of whether the margin was calculated using all of the new respondents (*Poll Margin*) or just the new respondents who chose either Biden or Trump (*Poll Margin (No DK)*). It is interesting that many new voters declined to choose either Biden or Trump when initially asked (*% Initial DK/Refusal*).[42]

	Lower Bound Size	Voter File 1 Size	Voter File 2 Size	% Not Voting in 2016	Vote Margin (Est.)	Poll Margin	Poll Margin (No DK)	% Initial DK/Refusal
Arizona	23%	39%	36%	30	Biden +11	Biden +27	Biden +28	12
Florida	15%	31%	30%	19	Trump +14	Biden +25	Biden +32	20
Georgia	17%	31%	31%	22	Biden + 23	Biden +4	Biden +5	14
Iowa	7%	24%	23%	15	Biden +4	Biden +17	Biden + 22	16
Kansas	14%	29%	28%	17	Biden +18	Biden +11	Biden +13	10
Michigan	13%	34%	27%	14	Biden +18	Biden +39	Biden +56	18
Montana	18%	30%	22%	13	Biden +2	Biden +6	Biden +8	10
Nevada	20%	51%	38%	20	Biden +2	Biden +24	Biden +26	12
North Carolina	14%	30%	29%	17	Biden +11	Biden +21	Biden +29	16
Pennsylvania	11%	30%	25%	21	Biden +13	Biden +15	Biden +19	19
Wisconsin	10%	33%	25%	12	Biden +10	Biden +30	Biden +37	13

Table 11. Estimates of the Proportion of "New" Voters in the Electorate and Margin for These Voters. Estimates of the proportion of new voters are based on calculations using certified vote and voter-file estimates. Margin for new voters is estimated in each state based on state-wide partisan results for 2016 and 2020. Poll results are based on the analysis of a set of polls provided to the Task Force. Poll results are weighted.

Considering the vote choice of these new 2020 voters reveals substantial support for Biden with a margin of support often far greater than the admittedly crude vote-based calculation would suggest. While this cannot prove that the polling error was related to the new voters, the magnitude of the effects reported in Table 11 could easily account for the polling error.

[42] Roughly half of these respondents gave an answer when pushed. The distribution was roughly similar to the initial break and the sample was small; nonetheless, this initial reluctance is mentioned and emphasized in case those reluctant voters were less likely to report truthfully.

In Arizona, for example, among the 30% of respondents who reported not having voted in 2016, the 28-point Biden margin changed the overall margin by 8.4 points (.3 x .28) toward Biden. (If we exclude the 12% who did not give a candidate response, the margin still moved 7.4 points toward Biden because of these voters (.88 x .3 x .28).) Even in states with a smaller proportion of new voters such as Wisconsin (12% of the sample) or Michigan (14%), the large pro-Biden margins among new voters in these states greatly affected the overall margin. In Michigan, the 56-point Biden margin among voters who reported voting in 2020 but not 2016 shifted the overall margin by 7.8 points (.14 x .56), and the 37-point Biden margin in Wisconsin shifted the overall margin by 4.4 points (.12 x .37).

Precise and conclusive statements are impossible given the number of assumptions required to produce the benchmark comparisons reported in Table 11. Even so, the discrepancies in terms of the size and vote choice of new voters are noteworthy. The polling margins are far more supportive of Biden than the vote-based approximations would suggest (except in Georgia and Kansas) and the magnitude of the differences (e.g., 39 points in Florida, 21 points in Michigan, 20 points in Wisconsin). These differences could have noticeable effects on the polling error given the voter-file estimates of how many new voters there may have been in 2020. If the margin among new voters in the polls was 20 points too favorable towards Biden among the 30% of new voters (according to the voter file estimates) this would produce six points of polling error (.30 x .20). Six points of polling error is larger than the average polling error documents in part one.

To be clear, the 2020 poll data do not allow the conclusion to be drawn that new voters who responded to polls were representative of all new voters, or that the polls had too many or too few new voters in their samples. Also, it is not sure how the new voters who responded to polls compare in number and opinion to those who did not. Different calculations and data sources provide different benchmarks. It is also unsure whether or not the patterns found in the polls can be generalized more broadly.

Even with all this uncertainly, clearly there were enough voters who voted in 2020 but not 2016 to account for a large portion of the polling error. Among the polls examined, the new voters seemed to be too few and too supportive of Biden relative to the benchmarks. The pattern is consistent with the polls missing new Trump voters, but without knowing the actual number and vote choice of the new voters and who did and did not respond to polls, definitive conclusions are impossible.

13. Error Due to Likely Voter Models, Noncoverage, or Nonresponse: Likely to Be Factors

Reweighting the results live phone (RBS and RDD) polls to match the 2020 outcome required the largest changes to the weights for partisanship and 2016 vote choice. For weighted poll results to match the 2020 outcome, the percentage of Republicans (or 2016 Trump voters) had to be increased by several points and the percentage of Democrats (or 2016 Clinton voters) had to be decreased by several points. Similar patterns were found for RBS and RDD surveys, suggesting that the polling error is not due to coverage issues related to the use of voter files. Implications pertaining to sources of polling error depend on the nature of the nonresponse. If the opinions of the partisans included in polls reflect the opinions of the partisans not included in the polls, then there is no within-group nonresponse and the patterns would identify the source of polling error as having too many Democratic respondents and too few Republicans. On the other hand, if Republicans (or independents or new voters) included in 2020 polls were more likely to support Biden than those not included in the polls, then it is impossible to identify why the polls differed from the final outcome based on the available data.

Some explanations of polling error are harder to prove than others because of lack of data and because the polling error is consistent with multiple observationally equivalent mechanisms.

The term “missingness” encompasses those voters excluded from the possibility of contact (e.g., not in a voter file list, contact information missing, or lack of internet access) as well as individuals contacted but unwilling to participate in polls. Similar patterns are observed across various sampling and contact methods, so polling errors seem more likely to be from lack of response than from lack of coverage; however, coverage issues cannot be ruled out entirely. Therefore the term “missingness” is preferred rather than “nonresponse.”
In general, surveys using random digit dialing (RDD) to contact respondents weight the data to match population parameters for the general public and then use questions about voter registration and likelihood of voting to identify the electorate. (Panels with a general population sample can also use a similar procedure.) In contrast to RDD, registration-based samples (RBS) must directly estimate the likely demographic composition of the electorate.

RBS, if possible, uses reliable variables such as sex, age, and region/location (and perhaps race in some states) from the voter file, but other adjustments are more difficult. Education, in particular, is problematic. It is not clear how well the available information (e.g., the Current Population Survey’s most recent voting supplement) can provide guidance on the educational distribution of the current electorate.

Considering these difficulties, the polling error could well have resulted from incorrect assumptions about the electorate, i.e., about who would vote in 2020. Pollsters routinely adjust their samples to account for those excluded from the sample, those unwilling to respond, and likely voters, but these adjustments may have predicted electorate that did not materialize. Perhaps they postulated an electorate that was too Democratic because of assumptions about the age, race, or even partisanship of the electorate. Details about statistical adjustments are rarely reported in sufficient detail to allow reviewers to ascertain what was done, hence it is difficult to fully assess the effects of such adjustments on polling error.

A second possibility is that the failed attempts of pollsters to account for missingness by not fully taking into account the nature of the missingness. They may have failed to fully account for unwillingness as a basis for nonresponse or to fully account for noncoverage through post-stratification, which seeks to adjust the weights of undersampled and oversampled groups.

Even with correct weighting targets, voters could be excluded or included in the sample because of the parameters that weight the data. That is exactly what caused some of the polling error in 2016. The sample of respondents included in polls was unbalanced in terms of education because respondents with higher education levels were more likely to participate in polls and vote for Democratic candidates. As a result, vote margins in the 2016 polls were more favorable for Democrats than the actual vote margin.

Many such factors could have contributed to polling errors in 2020 either individually or collectively. A non-exhaustive listing of plausible explanations related to nonresponse issues follows.

- The record-breaking turnout in 2020 included many new voters. Perhaps the new voters who responded to polls in 2020 differed from the new voters who did not respond to polls. Perhaps the new voters mobilized by Trump were less likely to cooperate with polls than new voters mobilized by Biden. (See Section 11.)
- Perhaps Republicans were less likely than Democrats to respond to polls in 2020 (because of the politicization of polling by Republican leaders) in ways that conventional weighting methods missed (Gelman et al 2012). Between-group differences in survey participation related to differences in vote choice could produce the types of error documented in this report.
- Even if the proportion of Republicans and Democrats who responded was correct and hence there was no between-group nonresponse, perhaps Republicans who responded were more likely than those who did not respond to support Democrats, i.e., perhaps there was within-group nonresponse.
- Perhaps there was both between-group and within-group nonresponse. For example, perhaps Republicans were less likely than Democrats to respond and those Republicans who did respond were more likely to support Democrats than those who did not respond.
- Similarly, perhaps independents were less likely than others to respond and those independents who did respond were more likely to support Democrats than those who did not respond.

Unfortunately, the ability to determine the cause or causes of polling error in 2020 is limited by the available data. Unless the composition of the overall electorate is known, looking only at who responded says nothing about who did not respond. Not knowing if the Republicans (or unaffiliated voters, or new voters) who responded to polls were more supportive of Biden than those who did not respond, for example, it is impossible to identify the primary source of polling error.

The number of potential causes can be narrowed through a comparison of weighted poll results with the reweighting necessary to match the certified vote by geography and method of voting. The composition of the electorate implied by the poll results as originally weighted differs from the composition of the electorate implied by poll results as reweighted to match the outcome of the election. Thus, reweighting provides clues to the factors that likely contributed to the poll error. The comparison reveals how sampling and weighting can be adjusted to match the poll to the outcome and helps identify characteristics most correlated with polling error.

Reweighting the poll results to match the 2020 election required increasing the percentage of self-identified Republicans and decreasing the percentage of Democrats, no matter whether RBS or RDD polls were examined at either the national or state level. While consistent with a polling error caused by an incorrect partisan distribution, it is important to note that correlations and causative factors are not identical. Causation should not be assumed because of correlation. Our work identified the characteristics most related to polling error, but there are many ways that the weights could be adjusted to match the final 2020 outcome. More importantly, these correlations found with partisanship could result from many different sources of polling error.

If the vote choices of a particular group included in polls mirror the vote choices of voters from the same group not included in the polls, then the cause of error is narrowed to having too many or too few respondents sharing a demographic characteristic, and our analysis would reveal which groups were over- and underrepresented in the originally weighted sample. If the largest differences occur among variables that were already being used in the statistical adjustments to correct for missingness, then the original weighting targets were likely incorrect.

For example, if the original weighting has 25% of the likely voters under the age of 24 and the outcome reweighting has only 15% of the reweighted sample under the age of 24, then this difference suggests the original weights relied on incorrect assumptions about young respondents in 2020. On the other hand, if the largest differences occur among demographics not used to weight poll results, then results could highlight the effects of using those demographics to weight the results. Conditional on knowing the final outcome, adjustments based on those variables would improve the accuracy of the poll.[43]

Within the same demographic group, if the opinions of those who responded differ from the opinions those who did not respond, then it is impossible to identify the cause or causes of polling error based on correlations. Only the opinions of those who responded are known. Reweighting cannot account for differences in the opinions between those who were and were not included in polls! (Perhaps the differences are what motivate some to respond.)

A hypothetical example helps illustrate the problem. Suppose the polling error was entirely caused by unaffiliated voters. In particular, suppose unaffiliated voters who supported Biden were eager to participate in polls so they could express their support for Biden over Trump, and suppose unaffiliated voters who supported Trump did not trust the media or polls and were unwilling to participate in polls when asked. This pattern of nonresponse would produce a pro-Biden polling error because the unaffiliated voters included in polls were more supportive of Biden than the unaffiliated voters in overall electorate. Even if the true source of polling error was not having enough Trump-supporting unaffiliated voters in the sample, reweighting the responses to match the 2020 outcome would show the largest changes in terms of the percentage of Republican and Democratic respondents even if they were correctly sized in the weighted poll results.

Our strategy of reweighting the originally weighted poll to match the 2020 certified outcome only identifies characteristics most associated with polling error *among those included in the poll*. Without knowing the opinions and size of unaffiliated voters who chose not to participate, it is impossible to rule out this possible explanation. Information about nonparticipants by definition does not exist. Our analysis identified the characteristics most correlated with the polling error among those included in the polls. Still, it is vital to note this important caveat: a correlation with polling error does not necessarily identify the cause of polling error.

The ways polls that polls are reweighted to change the electorate and match the 2020 outcome are examined in the following subsections, beginning with RBS polls in Section 13.1 and RDD polls in Section 13.2). Reweighting changes the implied electorate in terms of standard demographics and partisanship.[44] The reweightings that were performed always had the largest effect on the partisan composition of respondents.

Section 13.3 considers how the distribution of self-reported 2016 vote changes. Reweighting to match the 2020 outcome requires a much larger implied Trump-Clinton margin than actually occurred. Either 2016 Trump voters were much more likely to vote in 2020 than 2016 Clinton voters, or the 2016 Trump voters included in polls were more likely than those not included in the polls to vote for Democrats.

[43] Whether or not those variables should be used in the weighting is less clear. Even if a variable is closely related to the missingness in 2020, it does not necessarily follow that the variable should be used to account for missingness in circumstances where the truth is unknown. The usefulness of the variable for accounting for missingness prior to an election depends on defining the correct weighting targets, which may be a very difficult task.

[44] These analyses were conducted using 12 RBS polls and seven RDD polls. The Task Force is grateful to the pollsters who made these data available. These polls are only a subset of all of the polls conducted in 2020 and may not be representative of non-telephone polling or other telephone polls.

Finally, Section 13.4 demonstrates that the polls weighted using partisanship and 2016 voting behavior were no more accurate than polls not weighted in this manner.

To determine the effects of reweighting the polls to the actual 2020 election results, weighting targets were defined using the county-level certified vote totals for Trump and Biden, including the votes cast via absentee voting, in-person early voting, and in-person Election Day voting according to the voter files.[45] County-level counts were aggregated to county strata based on whether the county's past voting behavior was categorized as Strong Republican, Moderate Republican, Toss-Up, Moderate Democratic, or Strong Democratic.[46] Appendix D describes the reweighting process in more detail, but the important point is that the data was reweighted to match the certified outcome in terms of both geography and method of voting.

Strata-level targets for method of voting and candidate vote share were used to reweight polls. AAPOR Transparency Initiative members shared their individual-level poll results with the Task Force. One member organization shared RBS polls from AZ, FL, GA, IA, KS, MI, MT, NC, NV, PA, TX, and WI. Another member organization shared RDD polls from AZ, FL, MI, NC, PA, WI, and a national poll. Both sets of polls used live human interviewers but there were differences as follows: method of sample selection (RBS vs. RDD), likely-voter screen (the RBS poll treated every respondent as a likely voter and the RDD survey used an explicit screen based on poll questions), weighting adjustments, and question wording. Reweighting the RDD and RBS polls produced similar patterns, so the differences between RDD and RBS polls are unlikely to be primarily responsible for polling error.

13.1 Reweighting RBS Polls to Match the 2020 Outcome

> *Reweighting the weighted results of RBS polls to match the 2020 outcome suggests that the largest changes occurred in terms of partisanship and 2016 self-reported vote choice. To match the certified 2020 outcome, the samples were reweighted by increasing the percentage of self-reported 2016 Trump voters by two points and decreasing the percentage of self-reported 2016 Clinton voters by two points. The composition of the implied electorate in terms of the demographics used to weight the sample were largely unchanged.*

The results for every state are graphed in Appendix E, but it is useful to focus on a single state to describe the results and general pattern. Pennsylvania was chosen because it is about average in terms of signed polling error (3.6 percentage points too Democratic). The outcome in Pennsylvania was used to project Biden as the winner of the 2020 election.

Figure 16 graphs how reweighting affects the response categories listed along the left-hand-side of the plot. Positive values indicate a greater proportion of respondents in the outcome-reweighted data compared to the originally weighted data, and negative values indicate the original sample had too many with that response or the statistical adjustment overrepresented the size of that group.[47] Horizontal lines distinguish the questions.

[45] The Task Force thanks Rezwana Uddin for extracting this information.
[46] Arizona is handled differently given the small number of counties. In Arizona, the strata are based on Maricopa County, Moderate Democratic, and Moderate Republican.
[47] This is not the only interpretation; as previously emphasized, increasing or decreasing the size of a group to match the election outcome does not prove that the support of the group is responsible for the polling error, especially if within-group missingness occurs among any group of voters. The effects of adjustments may establish the correlation of the size of that group with the polling error, but correlation cannot establish causation.

Starting from the top of Figure 16, the proportion of respondents who indicate that they do not know whether they would self-identify as a Democrat or a Republican is largely unchanged (a fractional decrease) after reweighting the poll to match the 2020 outcome. The same applies too for "other" in response to partisanship. The change in the proportion who self-identify as an independent is slightly negative but close to zero, indicating fractionally fewer self-identified independents after reweighting the data to match the 2020 outcome. Much larger changes occurred among self-identified Republicans and Democrats. To make the weighted poll results match the 2020 outcome requires increasing the proportion of Republicans and decreasing the percentage of Democratic respondents by nearly two percentage points.

Table 12 shows the relationship between partisanship and polling error. It lists the percentages of respondents who self-identify with each party in the unweighted data and under the two poll weightings. Relative to the unweighted data, the original weights decreased self-identified Democrats by one percentage point and increased self-identified Republicans by one percentage point. Reweighting those weighted results to match the 2020 certified vote in Pennsylvania decreased self-identified Democrats by another two percentage points and increased self-identified Republicans by the same amount. Overall, a three percentage point shift relative to the unweighted sample and a six percentage point shift in the net partisan margin (from Democrat +5 to Republican +1) is required to match the poll responses to the 2020 outcome.

	Unweighted	Original Weight	Outcome Reweighted
Democrat	41%	40%	38%
Republican	36%	37%	39%
Independent	19%	19%	19%
Other	2%	2%	2%

Table 12. Partisan Composition of Pennsylvania RBS Poll. Because this was a RBS poll, the unweighted partisan composition may reflect partisan targets in sample selection.

Substantial changes result from adjustments for partisanship and self-reported 2016 presidential vote choice. Changes to other variables commonly used to adjust poll error are also worth noting; however, the composition of the electorate in terms of age, race, gender, and education are relatively unchanged after reweighting the poll to match the 2020 outcome. For example, the polling error would decrease if the proportion of respondents with no more than a high school education were increased by as little as one half of a percentage point (offset by a decreased proportion with a graduate degree) but the difference would be small relative to the polling error and relative to the changes implied in the partisan composition.

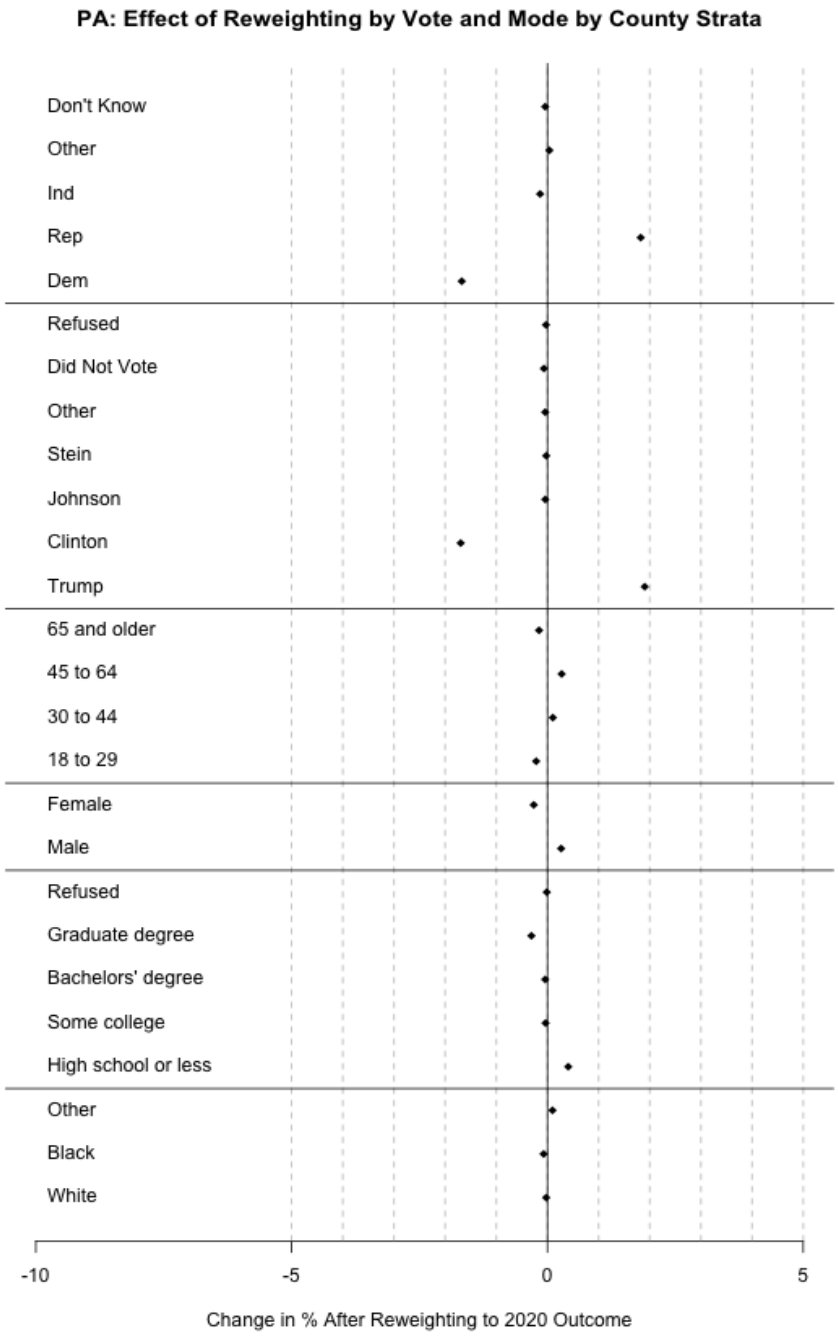


Figure 16. Comparing the Composition of the Pennsylvania Electorate Using the Original Weights and the 2020 Outcome Weights. The outcome-adjusted weight is a result of reweighting the likely voters in the original poll to match the method of voting and the proportion of votes for each candidate by county party stratum.

Stepping back, reweighting the Pennsylvania RBS poll reveals two important patterns. First, there were only minor changes in the proportion of respondents belonging to each response category among standard demographic questions. This suggests, but does not prove, that the polling error is not easily corrected using standard demographic adjustments. Second, large differences exist in terms of partisanship and self-reported 2016 vote. As previously noted, such differences could be caused by many potential sources of polling error.

The pattern found in Pennsylvania was true for every poll examined. In general, the weights fr standard demographics were largely unchanged after reweighting the weighted poll to match the 2020 outcome. Figure 17 illustrates this finding by graphing the change in the percentage of respondents of varying education levels that results from reweighting each poll to match the 2020 outcome. In general, most differences are smaller than one percentage point and the larger differences (e.g., in Montana, the proportion of those with bachelor's degrees would have to be decreased by two percentage points and the percentage of those with some college and high school or less would have to be increased by a percentage point each) are small relative to the magnitude of the signed error (e.g., the average signed error in Montana was 10.7 points). Overall, the polling error was only weakly related to standard demographics.

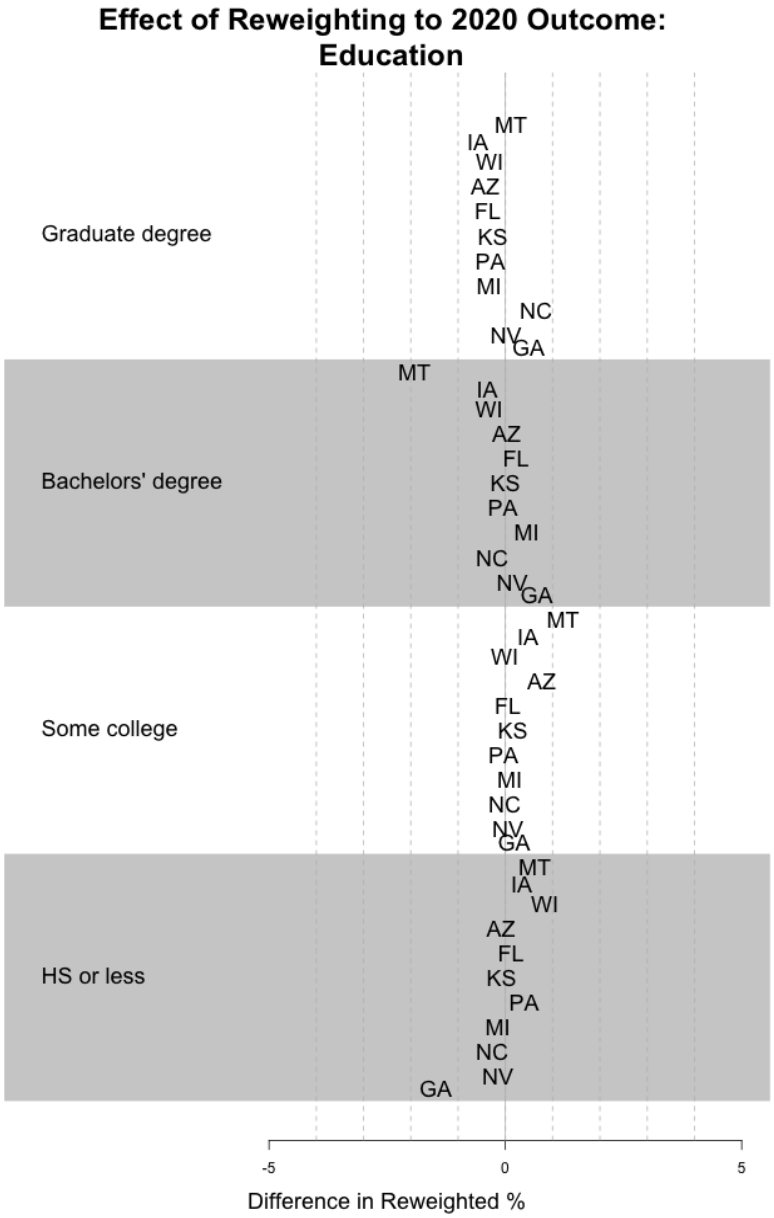


Figure 17. The Change in Electoral Composition for Education. This plot shows how the estimated proportion of the electorate at each education level differs between the original weights and the outcome-adjusted weights. The polls analyzed are RBS telephone polls.

Much larger differences emerge for partisanship and self-reported 2016 vote. Compared to the original survey weights, reweighting the 2020 survey data to match the 2020 outcome produced an implied electorate with more Republicans and fewer Democrats, as well as more 2016 Trump voters and fewer 2016 Clinton voters. Figure 18 shows that on average, in every state except for Georgia (which also was among the states with the smallest polling error), the percentage of self-reported 2016 Trump voters would have to be increased by nearly two percentage points (and the percentage of 2016 Clinton voters would have to be decreased by a similar amount) for the weighted survey data match the 2020 outcome.

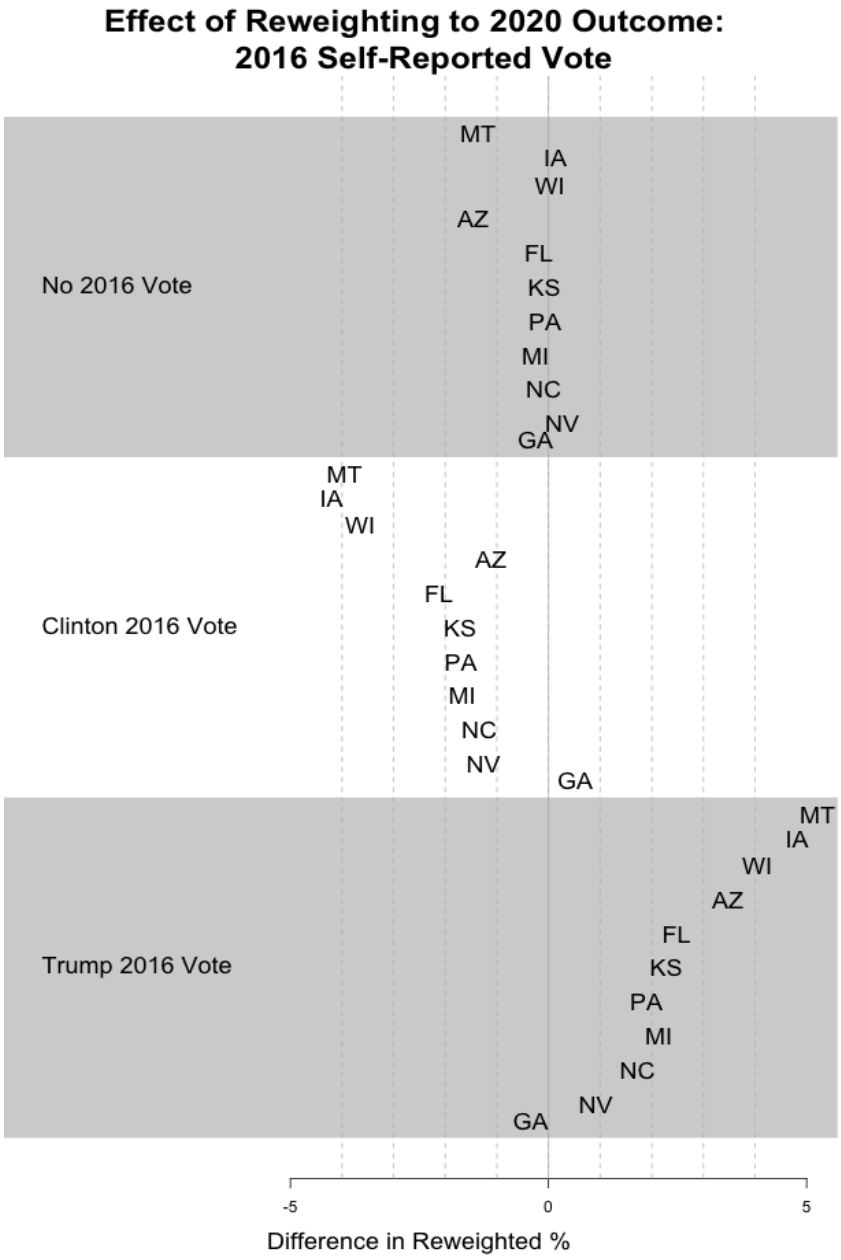


Figure 18. The Change in Electoral Composition for 2016 Vote. This shows how the estimated proportion of the electorate giving each response differs between the original weights and the outcome-adjusted weights. The polls analyzed are RBS telephone polls.

13.2 Reweighting RDD Polls to Match the 2020 Outcome

Reweighting the weighted results of RDD polls to match the 2020 election outcome produced the largest changes in the distribution of partisanship for both state and national polls. To match the certified 2020 outcome required increasing the total percentage of self-reported Republicans by around four points and decreasing the total percentage of self-reported Democrats in the reweighted sample by around four points relative to the percentage in the originally weighted sample. Other demographics remained unchanged.

Reweighting the RDD phone polls revealed a similar pattern as reweighting the RBS polls. After the survey results were reweighted to match the 2020 outcome, there were minimal differences in the percentage of respondents belonging to the various response categories for demographics traditionally used to post-stratify. The largest and most consequential differences occurred in the distribution of self-reported partisanship. Figure 19 summarizes the effect for partisanship. To match the poll with the outcome, the combined proportion of strong and leaning Republican respondents would have to be increased by roughly four percentage points in the sample, and the combined proportion of strong and leaning Democrats would have to be decreased by a similar amount.

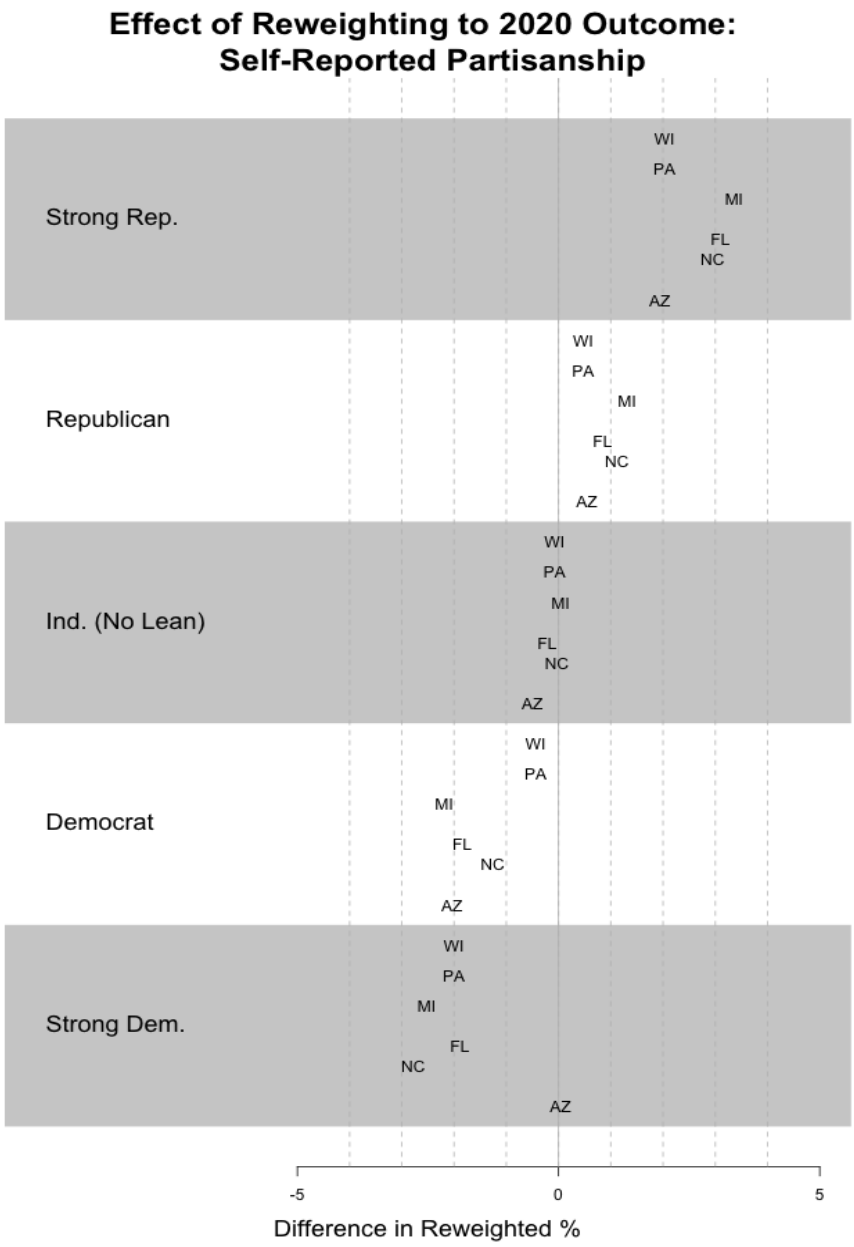


Figure 19. The Change in Electoral Composition for Self-Reported Partisanship. This shows how the estimated proportion of the electorate giving each response differs between the original weights and the outcome adjusted weights. The polls analyzed are RDD telephone polls.

A similar analysis using a national RDD survey yields a similar conclusion. The national RDD poll by size and candidate vote were reweighted within county quintiles. County-level vote results provided by the National Election Pool were sorted based on the margin between Biden and Trump within each county to determine quintiles representing roughly 20% of the total nationwide vote.[48] The original poll weights were adjusted so results among likely voters within each quintile matched the correct size and distribution of vote.

[48] Election results are not available by county in Alaska or Washington, D.C. The jurisdiction-wide results were treated as county-level results in creating the quintiles for those places.

Table 13 shows that the original poll results were relatively close to the correct sizes of the five quintiles. The largest difference is a slight *overstatement* of the most pro-Trump counties where the original weighting results in a likely vote size of 23% of the electorate compared to the 20% actual vote. However, the unreported estimated vote shares was significantly off within these quintiles, particularly in the most pro-Trump quintile. Consistent with the patterns of part one, these effects of differential nonresponse were deepest in the most Republican counties nationwide, even as those counties appeared well-represented in terms of size.

	Size in Actual Vote	LV Size Using Original Weights
Quintile 1 – Strongest Trump	20%	23%
Quintile 2	21%	20%
Quintile 3	20%	18%
Quintile 4	20%	19%
Quintile 5 – Strongest Biden	20%	20%

Table 13. Comparing the Size of County Quintiles Using the Original and Outcome-Reweighting. Results are from a national RDD poll.

As in the state-level reweighting analysis, reweighting the national RDD poll to match the outcome found little or no movement in measures not related to partisanship or candidate support. Table 14 shows that none of the standard demographics changed in size by more than a single point, and only partisanship and ideology were affected by the reweighting. To make the weighted poll results match the final outcome required increasing the percentage of Republicans in the sample by 3.1 percentage points and the percentage of Republican leaners by 0.9 percentage points while also reducing the percentage of Democratic respondents by 2.3 percentage points and the percentage of Democratic leaners by 1.2 percentage points.

	Original Weight	Outcome Adjusted Weight	Change in % After Weighting
Republican	30.7	33.7	+3.1
Lean Republican	12.5	13.4	+0.9
Independent (no lean)	4.1	3.6	-0.5
Lean Democratic	17.0	15.8	-1.2
Democratic	35.7	33.5	-2.3
Don't Know	3.2	2.5	-0.6
Conservative	32.4	34.7	+2.3
Moderate	41.2	40.6	-0.5
Liberal	23.3	22.2	-1.2
Don't Know/Refused	0.7	0.6	-0.1
65+	24.4	24.6	+0.2
50-64	27.3	27.5	+0.3
30-49	31.4	31.2	-0.2
18-29	16.3	16.1	-0.2
Female	52.6	52.2	-0.5
Male	47.4	47.8	+0.5
Don't Know/Refused	8.6	8.5	-0.1
$100K+	30.6	31.0	+0.4
$50K-<$100K	29.9	30.3	+0.4
<$50K	30.9	30.3	-0.6
Don't Know/Refused	1.0	0.8	-0.1
College grad	39.1	39.1	-0.02
Some college	28.9	28.9	+0.1
HS Grad	27.9	27.9	+0.1
No HS	3.2	3.2	-0.02
Don't Know/Refused	1.4	1.3	-0.1
NET People of color	31.4	30.7	-0.7
Latino	11.3	11.2	-0.01
Black	11.7	10.9	-0.7
White	67.2	68.0	+0.8

Table 14. The Change in the Size of Demographic Groups Using Original Weights and Reweighting to the Outcome. The results are for a national RDD poll.

13.3 Within-Group Differences in Who Responded to Polls

> *Reweighting polls to match the 2020 outcome required that the Clinton-Trump margin among self-reported 2016 voters be far more favorable to Trump than the actual 2016 margin. As a result, using the self-reported 2016 vote to weight to the past vote does not fix the polling error. If the 2016 voters who voted in 2020 were representative of the 2016 electorate, then 2016 Trump voters who participated in polls would be more likely to vote for Biden than those who did not participate. However, if the 2016 voters who voted in 2020 were not representative of the 2016 electorate, then it would be impossible to draw conclusions: it would be unclear what the implied 2016 margin should be among the 2016 voters who voted in 2020. That is another complication of weighting the results using 2016 past vote.*

The results of Sections 13.1 and 13.2 are consistent with polling error caused by having too many Democrats and not enough Republicans in the sample. However, this interpretation depends critically on the assumption that the individuals who responded were representative of those who did not (i.e., it is conditional on the measured characteristics).

As previously noted, it difficult to interpret the findings of Sections 13.1 and 13.2 as identifying the source of the polling error if the voters who responded to polls in 2020 may have differed from those who did not even within a demographic category. The importance of this subtle point is perhaps best illustrated by considering the results of the new voter analysis in Section 12 alongside the results of Section 13.1. Section 12 showed that perhaps too few new voters were included in the polls and those new voters may also have been too supportive of Biden compared to new voters overall. But the analysis of those same polls in Section 13.1 showed that the percentage of 2020 voters who reported not voting in 2016 did not change after reweighting to match the 2020 certified outcome. Does this mean that new voters were not responsible for the polling error?

Within-group differences in who responds to polls make it impossible to interpret the results of Sections 13.1 and 13.2 as revealing the cause of polling error. The analysis determines the characteristic most related to the polling error among those who respond; if the new voters who responded to the polls were more supportive of Biden than those who did not respond to polls, and if this differential nonresponse was the true cause of the Biden-Trump overstatement in the polls, then reweighting the polls to match the 2020 outcome would produce the exact pattern found in Sections 13.1 and 13.2. Even if the polling error was caused by new voters, the outcome-based reweighting would up-weight Republicans and down-weight Democrats because the characteristic most related to the polling error *among those who responded* is partisanship.

A similar argument can be made for unaffiliated voters, Republican voters, or any group in which vote choice and willingness to participate could be related. Only if we are willing to assume that the individuals who responded were representative of those who do not (conditional on the characteristics being used to account for nonresponse) can we interpret the findings of Sections 13.1 and 13.2 as evidence of the source of the polling error being too heavily Democratic. If there are differences in who responds to polls within a group, as in the hypothetical case of new voters just discussed, then it becomes far more difficult to identify the cause (and potential correction). But without information on the opinions of those missing from polls, it is impossible to conclusively answer whether those who responded to polls differed from those who did not.

Suggestive hints are possible if we are willing to assume 2016 voters were equally likely to vote in 2020 regardless of who they voted for in 2016. If this assumption is true, then the margin of self-reported 2016 vote among the

2016 voters who voted in 2020 should closely match the 2016 vote certified margin. (Polls that used self-reported 2016 voting behavior must make this assumption to weight their results to match the 2016 certified vote.) Consequently, the margin of the 2016 self-reported vote among the outcome-reweighted polls can be used to evaluate whether the 2016 margin matches the implied margin in the reweighted data. If it doesn't match, then the 2016 voters who responded to polls in 2020 may have differed from those who did not.

If the assumption is not allowed that the 2016 voters who voted in 2020 were representative of the 2016 electorate, then the margin of self-reported 2016 vote would not be expected to match the 2016 certified margin and nothing could be learned from the comparisons. If the 2016 voters who voted in 2020 are not representative of the actual 2016 outcome, then the distribution of their self-reported vote would match the actual 2016 vote. No benchmark would be available for comparing the respondents in 2020. Attempts to weight poll results in 2020 to match the 2016 vote using self-reported past vote would also be erroneous because there would be no reason to expect the distribution of self-reported past vote to match the actual 2016 certified vote.

Table 15 lists for 11 states the 2016 certified vote margin, the margin estimated by the 2016 self-reported vote question in the original unweighted poll data, the margin after using the original survey weights, and the margin when using the outcome adjusted reweighting.[49] The final column compares how much the margin between Trump and Clinton differs between the unweighted and the outcome-reweighted results. Several patterns emerge.

	2016 Certified Margin	Unweighted 2016 Margin	Original Weighted 2016 Margin	Outcome Reweighted 2016 Margin	Total Shift
Arizona	Trump +4	Trump +5	Trump +7	Trump +13	Trump +8
Florida	Trump +1	Clinton +1	Trump +4	Trump +10	Trump +11
Georgia	Trump +6	Clinton +1	Trump +3	Trump +3	Trump +4
Iowa	Trump +10	Trump +7	Trump +11	Trump +22	Trump +15
Kansas	Trump + 21	Trump +12	Trump +20	Trump +25	Trump +13
Michigan	Trump +0.3	Clinton +4	Trump +2	Trump +6	Trump +10
Montana	Trump +21	Clinton +8	Trump +3	Trump +13	Trump +21
Nevada	Clinton +2	Clinton +3	Trump +2	Trump +5	Trump +8
North Carolina	Trump +4	Clinton +1	Trump +6	Trump +10	Trump +11
Pennsylvania	Trump +1	Clinton +1	Trump +4	Trump +8	Trump +9
Wisconsin	Trump +1	Clinton +5	Trump +3	Trump +12	Trump +17

Table 15. Comparing the Effects of Reweighting 2020 Polls to Match 2020 Actual Vote on the Estimated 2016 Presidential Vote Share in the Polls. *Total Shift* is the difference between the unweighted 2016 margin among self-reported votes and the outcome re-weighted results.

[49] We exclude respondents who refused to answer or who indicated that they did not vote.

First, although the 2016 margins in the unweighted data are generally too Democratic (except for Arizona and Nevada), the original poll weights produced margins that closely matched the 2016 certified margin in every state except Montana. This means that the originally weighted sample used to estimate the 2020 election closely matched what happened in 2016. This point highlights that weighting to the 2016 past vote in 2020 does not eliminate the polling error. The polling error occurs even though the self-reported Clinton-Trump margin among 2016 voters closely matched the 2016 certified margin.

Second, reweighting the weighted results to match the 2020 outcome requires an electorate that supported Trump far more than the actual 2016 margin. In Wisconsin, for example, applying the outcome-adjusted weights produces a 12-point Trump margin among those who reported voting in 2016, a margin 11 points larger than the actual one-point 2016 Trump margin, and less accurate than the margin in either the unweighted data (Clinton +5) or originally weighted data (Trump +3).

The last column emphasizes just how much the implied 2016 vote share among those who reported voting in 2016 would have to change to make the 2020 poll results match the actual 2020 outcome. Except for Georgia, the adjustments required to match the 2020 vote outcome would increase the 2016 margin by double digits in favor of Trump. In Wisconsin, for example, reweighting the results to match the 2020 outcome required a 17-point margin for Trump among those who report voting in 2016.

Assuming the vote choice of 2016 voters in 2020 should match the 2016 certified outcome requires that the outcome-reweighted polls have a much larger margin for Trump than the actual 2016 outcome. That larger margin could be interpreted in several ways. It could mean that the 2016 Trump voters who responded were more likely to support Biden than the those who did not respond, or it could mean that the 2016 Trump voters included in polls were more likely to support Biden than those who were not included in the polls. In either case, reweighting the poll to match the 2020 outcome would require upweighting self-reported 2016 Trump voters to produce a margin for Trump far greater than the actual 2016 margin.

Assuming that the 2016 certified margin should be observed among 2016 voters voting in 2020 does not prove that the error was caused by 2016 Trump voters. Correlation does not prove causation: the same pattern could occur if the polling error were caused by differences in who responded among any group of respondents, e.g., among those who did not vote in 2016. The polling error is most related to 2016 vote choice, but it could emerge in many other ways that would produce the same pattern in the reweighted data. Without knowing how those included compare to those not included in the polls, it is impossible to identify the cause of the polling error.

If the self-reported past vote of 2016 voters cannot be expected to match the 2016 certified vote, then little could be inferred from Table 15 because there would be no benchmarks for comparing the opinions of the respondents. The implied 2016 margin for Trump may simply reflect that 2016 Trump voters were more likely to vote in 2020 than 2016 Clinton voters.

Without information on the size and opinions of those voters who were not polled, it is difficult to reach definitive conclusions about how issues of missingness may have affected polling in 2020.

13.4 Implications for Weighting by Partisanship and Past Vote

Polls weighted by partisanship and 2016 past vote were not more accurate, on average, than those weighted using only standard demographics. Weighting on partisanship and past vote may have reduced polling error, but it did not solve the issue.

If polling error in 2020 was caused by the unwillingness of Republicans or Trump-supporters to participate in polls, then did polls that used partisanship and 2016 vote share to adjust their responses produce more accurate estimates?

As part of its data collection effort, the Task Force attempted to obtain information on the variables used to statistically adjust every public poll available. This information was used to compare the average polling error of these polls, depending on whether polls were adjusted using partisanship or 2016 past vote. Among the 438 state-level presidential polls conducted in the last two weeks, 46 reported weighting on partisanship, 22 reported weighting on past 2016 vote, 249 reported a weighting scheme that used neither, and it was unclear as to how the sample was weighted for 121 polls.

Figure 20 reports the distribution of signed and absolute error depending on the parameters that used in the statistical adjustment. There are no obvious differences.

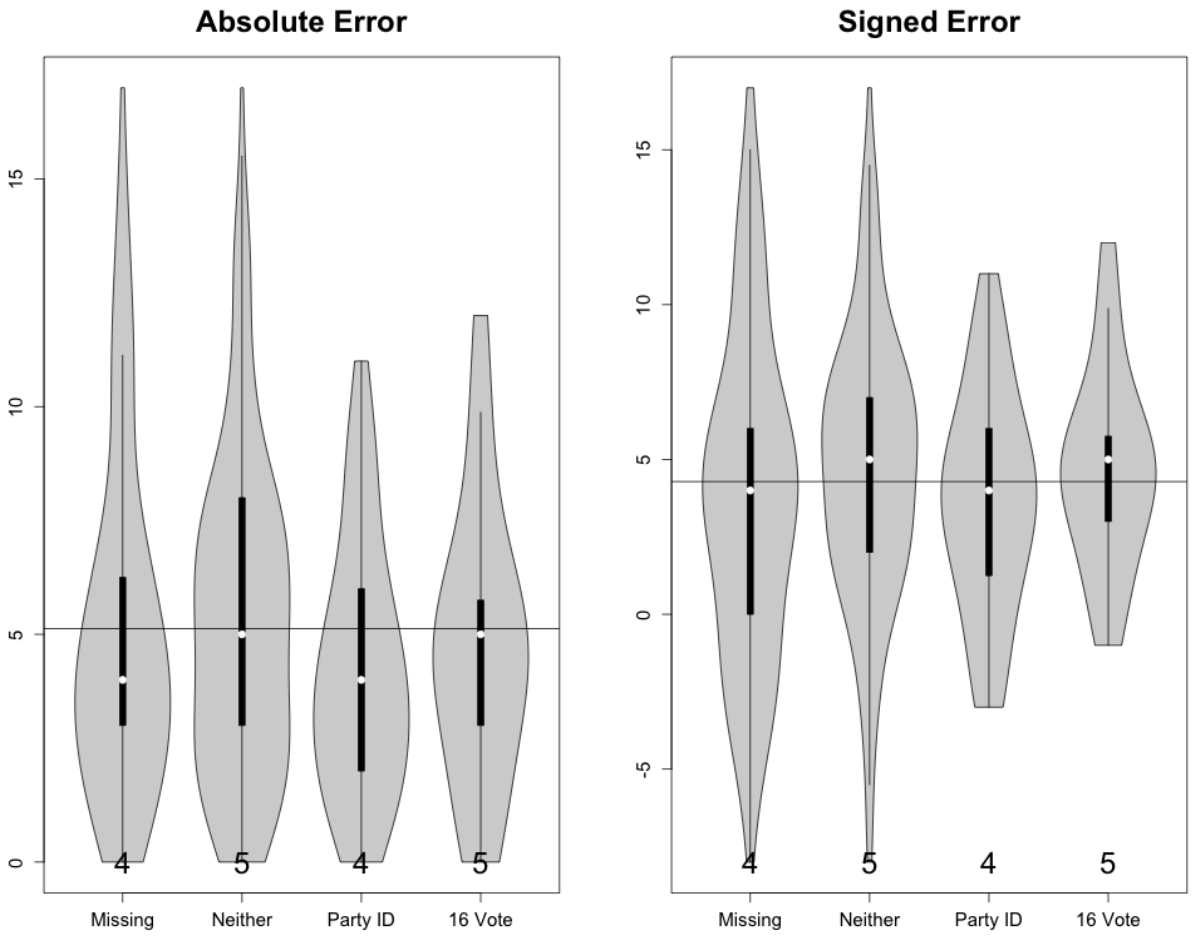


Figure 20. Average Error by Weighting Method. The number at the bottom is the median error for each set of polls. Among the 438 state-level presidential polls conducted in the last two weeks, 46 used *Party ID*, 22 used *2016 Vote*, 249 used *Neither*, and 121 were unclear about sample adjustment (*Missing*).

Even if polls were affected by nonresponse related to partisanship or 2016 Trump vote, polls using partisanship and past vote in their weighting were no more accurate. The similar level of average error may be due to incorrect assumptions about the proportion of partisans voting; 2016 voters who voted in 2020 not being representative of the 2016 electorate, which means weighting to the 2016 vote share being incorrect or partisans polled voting differently from partisans not polled. If the Republicans who participated in polls were more likely to support Biden than those who did not participate in polls, for example, weighting on partisanship would not eliminate the problem that is caused by the views of Republicans missing from the sample. While it is possible that the polls that chose to weight using partisanship and past vote would have performed worse had they not done so, weighting on partisanship and past vote clearly does not provide an easy solution.

14. Conclusions

Considering that whether Trump or Biden won the Electoral College in 2020 was decided by fewer than 50,000 voters in three states, pre-election polling faces many challenges. In close races, small changes in accuracy can have large consequences on predictions of winners, especially when voting is closely related to partisanship. If a poll overstates the percentage of Democrats by two percentage points, then a four-point Biden margin (say 50%-46%) appears in what would otherwise be a tied race (48-48). Errors can develop from statistical adjustments or issues related to noncoverage or nonresponse. One error can have a large effect and many small errors can add up to even larger errors, especially when margins are close.

Polls are often misinterpreted as precise predictions. It is important in pre-election polling to emphasize the uncertainty by contextualizing poll results relative to their precision. Considering that the average margin of error among the state-level presidential polls in 2020 was 3.9 points, that means candidate margins smaller than 7.8 points would be difficult to statistically distinguish from zero using conventional levels of statistical significance. Furthermore, accounting for uncertainty of statistical adjustments and other factors, the total survey error would be even larger.

The temptation to interpret polls as precise predictions arose in part because polls did so well predicting outcomes in 2008 and 2012. Nonetheless, putting poll results in their proper context is essential; whether or not the margins are large enough to distinguish between different outcomes, they should be reported along with the poll results. Most pre-election polls lack the precision necessary to predict the outcome of semi-close contests. Despite the desire to use polls to determine results in a close race, the precision of polls is often far less than the precision that is assumed by poll consumers. The polls themselves could exacerbate problems of scientific legitimacy when they are unable to live up to unrealistic expectations.

Our data collection and analyses provide a comprehensive characterization of polling error across races and states in the 2020 election. Our investigation reveals a systemic overstatement of the Democratic-Republican margin in nearly every contest, regardless of mode or proximity to the election. This overstatement is largest in states with more Republican supporters. Comparing the topline poll results to the certified vote share suggests that polls understated Trump support and overstated Biden support. The magnitude of those effects is dependent upon whether the reported candidate marginals or the two-candidate support was used.

Admittedly, this report is unsatisfyingly inconclusive as to why the polling error in 2020 was the largest in at least 20 years. Although some explanations can be ruled out, knowing what really happened is impeded by not knowing how those who participated in the polls compare to those who did not. Nothing can be learned about those who did not respond by examining those who did respond. Multiple explanations can be proposed for every pattern found, yet it cannot be conclusively determined which are valid.

Conclusions that can be drawn from our investigation are as follows.

It seems plausible that many issues were caused by nonresponse. Nevertheless, it is so far impossible to know the primary issue. Among the possibilities are: too many Democrats and too few Republicans responding to the polls (between-party nonresponse); the Democrats/Republicans who responded had different opinions than those who did not (within-party nonresponse); and new voters and independents unpredictable in terms of both size (too many or too few) and representativeness (i.e., were the new voters who responded similar to those who did not?). Any one of these, and possibly more than one, could produce the polling error. The primary source of polling error cannot be identified conclusively without knowing how nonrespondents and respondents compare.

It is unclear whether the problems polls faced in 2020 will persist in 2022 or 2024, and what happens in 2022 may be uninformative for knowing if there are longer-term issues. Because both the electorate and the set of respondents could change in 2022 relative to 2020 for multiple reasons, it is impossible to predict how the polls in 2022 might perform based on 2020 results.

There are at least three scenarios as follows.

Perhaps future pre-election polling will continue to be affected by the issues that affected the 2020 polls.[50] If the issues that affected the voters and respondents in 2020 persist without being corrected, the issues that affected polling in 2020 will also affect future polls. One plausible way that this would occur is if Republican-supporting voters have come to believe that polling is "fake news" and a tool of "voter suppression" so that the decision to not participate in a poll becomes a political act. Given the politicization of polling by Republican party leaders, it may be that the people who choose to respond or not differ in their opinions even conditional on observable demographics. This problem would be a difficult for pollsters to overcome.

Perhaps the problems of 2020 were unique to presidential elections because the polling error was caused by voters who vote in presidential elections but not midterm elections. If so, polls would be fine in 2022, but not 2024. Presidential elections are more salient and generate larger (and different) turnout than midterm elections, hence the polling error may matter most in presidential election years. Presidential elections typically attract less engaged voters. If the polling error is caused by voters who only vote in presidential years, then how well the polls do in 2022 will not reveal if the problems have been solved. A good performance by polls in the 2022 midterm will not provide evidence that the issue with polling error is "solved" if the problem in 2020 was brought about by voters more likely to vote only in presidential elections. To this point, recall that the 2018 midterm polls had an average signed error of only 0.1 and an average absolute error that was half a percentage point better than the 2016 state-level presidential polls.[51]

Perhaps the problems of 2020 were unique to the politics of Trump or the pandemic so the polls could be fine in 2022 and 2024. If voters who were normally unmotivated to vote and participate in polls were motivated to vote specifically for Trump (but not participate in polls), then this could have created the polling error. If so, Trump's absence in future elections may lessen the polling error caused by these voters, assuming another candidate does not mobilize the same constituency. The fact that we did not see similar error in the 2018 midterm elections is consistent with a Trump-caused polling error. Alternatively, perhaps the poor performance of polls in 2020 relative to 2018 reflects the effects of the pandemic. Factors relating to the pandemic include uncertainty associated with voting and differences in how likely individuals were to respond to polls given partisan-related differences in how likely individuals were to be reachable at home (e.g., Clinton et al. 2021). If the polling error is due to Trump or the pandemic, then polls may perform better once those forces are no longer relevant.

[50] As an example, the statewide polls in 2016 also overstated the Democratic-Republican margin. It was thought that this gap was related to education and that weighing by education would solve the issue. In analyses conducted by the 2016 Task Force, education weighting did reduce poll error but did not fully solve the issue; the 2016 Task Force report makes clear that the 14-point error in a Michigan poll was decreased to a seven point error, which is similar in size to the error observed in 2020.

[51] The following excerpt is from the 2018 AAPOR Task Force report (unpublished): "Looking at *signed error*, the 2018 state polls averaged an error of .09 percentage point, the lowest error by this measure compared to the NCPP measure in 2014 and to the AAPOR measure for state presidential polls in 2000, 2004, 2012, and 2016. The AAPOR measure for 2008 was very close, -0.10. Looking at absolute error, the 2018 polls had an average error of 4.64 percentage points, somewhat better than the 5.1 in the AAPOR analysis of 2016 state presidential polls, but worse than the numbers found by AAPOR for state presidential polls in 2004, 2008, and 2012. The 2018 finding was a bit worse than the 4.06 average found for state U.S. Senate and governors polls in 2014."

Despite this ambiguity, several more points are worth highlighting for those who conduct and cover pre-election polls. These points are not unique to polling in 2020, but the performance of polls in 2020 reveals their continued importance for understanding what polls can and cannot do.

Polling results depend heavily on statistical adjustments and it is often unclear what is assumed and how sensitive the reported results are to the assumptions. The results of pre-election polling are increasingly as much a function of statistical adjustments, such as assumptions about likely voters, as they are a function of actual data. In many cases, how the results are adjusted and how sensitive the results are to particular choices and decisions is unclear (e.g., Cohn 2016). Poll consumers may be misled by false confidence in results unless they understand how alternative adjustments could change the results. Considering the consequences of these decisions, it is vital to describe them better and quantify their effects on poll results.

Weighting by partisanship or past vote does not obviously solve the problem. Adjustments based on partisanship and past vote might not solve the errors. The weighting targets for partisanship are often unclear, and the close connection between partisanship and vote choice in the contemporary political environment means that incorrect targets can greatly affect the accuracy of polls. Using past election results also cannot solve this issue because it is impossible to compare current voters with past-election voters who may have different opinions. Similarly, partisans voters who respond to surveys have the same opinions as partisans voters who do not respond (i.e., within-group missingness), which presents another challenge for polling.

The margin of error is not the same as polling error. Sampling variability is only one source of error. Many have emphasized this important point (e.g., Mercer 2016; Rothschild and Goel 2016). Focusing only on the margin of error of a proportion runs the risk of overstating the precision. Setting aside the statistical details of the computation of the margin of error on an estimated difference, even in the best circumstances, poll results often are discussed in a way that provides a misleading characterization of polling performance and accuracy.

It is difficult to interpret the accuracy of individual polls: individual polls may have low error for arbitrary reasons and their current good performance may not necessarily indicate future success. In 2020, for example, polls that simply increased the percentage who supported Republican candidates and polls weighted by spurious factors that increased the implied Republican support may indeed have been more accurate in 2020, despite such arbitrary adjustments. Both the accuracy of a poll and justifications for those adjustments are important to consider. Procedures that are not grounded in statistical theory or theories about the data generating process may employ arbitrary adjustments. Such procedures may produce seemingly “accurate” results that fail to generalize to other elections. Absent transparency on the adjustments being done, it is difficult to evaluate the performance of individual pollsters.

WORKS CITED

Ansolabehere, Stephen and Eitan Hersh. 2014. “Voter Registration: The Process and Quality of Lists." In *The Measure of American Elections*, eds. Barry C. Burden and Charles Stewart III. Cambridge: Cambridge University Press.

Barber, Michael, and Jeremy C. Pope. 2018. “Does Party Trump Ideology? Disentangling Party and Ideology in America” *American Political Science Review* 113(1): 38-54.

Broockman, David E, and Daniel M Butler. 2017. “The Causal Effects of Elite Position Taking on Voter Attitudes: Field Experiments with Elite Communication,” *American Journal of Political Science*, 61(1): 208 – 221.

Bullock, John G. 2011. “Elite Influence on Public Opinion in an Informed Electorate,” *American Political Science Review*, 105(3): 496 – 515.

Clinton, Joshua D., Jon Cohen, John Lapinski, and Mark Trussler. 2021. “Partisan Pandemic: How Partisanship and Public Health Concerns Affect Individuals’ Social Mobility During COVID-19,” *Science Advances* 7(2): DOI: 10.1126/sciadv.abd7204

Cramer, Katherine J. 2016. *The Politics of Resentment: Rural Consciousness in Wisconsin and the Rise of Scott Walker*. Chicago: University of Chicago Press.

Byler, D. 2020. ``Did `Shy Trump Voters' throw off the polls? Maybe not. *The Washington Post*. Accessed at: https://www.washingtonpost.com/opinions/2020/11/09/\did-shy-trump-voters-throw-off-polls-maybe-not/.

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

Cohn, N. (2020). “Are the Polls Missing Republican Voters?” *The New York Times*. Accessed at: https://www.nytimes.com/2020/07/29/\upshot/polls-political-party-republicans.html.

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

Franklin, Charles H. 2007. “The ‘Margin of Error’ for Differences in Polls,” University of Wisconsin Working Paper.

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

Habberman, Maggie. 2020. “To Trump, ‘the Polls That Matter’ Point to Victory. The Rest Are ‘Fake.’” New York Times. November 1. Accessed at: https://www.nytimes.com/2020/11/01/us/politics/trump-polls.html

Kennedy, C., N. Hatley, S, Keeter, A. Mercer, R. Igielnik, and F. Traylor. 2018. "Comparing Survey Sampling Strategies: Random-Digit Dial vs. Voter Files." Accessed at: https://www.pewresearch.org/methods/2018/10/09/comparing-survey-sampling-strategies-random-digit-dial-vs-voter-files/

Kennedy, C., and H. Hartig. 2019. "Response rates in telephone surveys have resumed their decline." Accessed at: https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/

Keeter, S. (2018). The impact of survey nonresponse on survey accuracy. In *The Palgrave Handbook of Survey Research*, pages 373-381. Springer.

Keeter, S., Kennedy, C., and Deane, C. 2020. "Understanding how 2020 election polls performed and what it might mean for other kinds of survey work." Pew Research Center. Accessed at: https://www.pewresearch.org/fact-tank/2020/11/13/understanding-how-2020s-election-polls-performed-and-what-it-might-mean-for-other-kinds-of-survey-work/.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley & Sons.

Lenz, Gabriel S. 2012. *Follow the Leader? How Voters Respond to Politicians' Policies and Performance*. Chicago, IL: University of Chicago Press.

Matthews, D. 2020. "Election results: Why the polls got it wrong." *Vox*. Accessed at: https://www.vox.com/policy-and-politics/2020/11/10/21551766/election-polls-results-wrong-david-shor.

Mercer, Andrew. 2016. "5 Key Things to know about the margin of error in election polls," Pew Research Center. Accessed at: https://www.pewresearch.org/fact-tank/2016/09/08/understanding-the-margin-of-error-in-election-polls/

Mitofsky , Warren J. 1998. "Review: Was 1996 a Worse Year for Polls Than 1948?," *The Public Opinion Quarterly*, 62:2 (230-249).

Page, L. 2020. "US election polls miss shy Trump voters and failed to predict the way votes were cast," Polls missed Trump voters and were worse than 2016". Australian Broadcast Corporation. Accessed at: https://www.abc.net.au/news/2020-11-09/\us-election-polls-misjudged-\voter-behaviour-trump-biden/12862288.

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.

Scot Schraufnagel, Michael J. Pomante II, and Quan Li. 2021. "Cost of Voting in the American States," in *Election Law Journal: Rules, Politics, and Policy*. Dec:503-509. http://doi.org/10.1089/elj.2020.0666

Scott, Alastair J. and George A. F. Seber. 1983. "Difference of Proportions from the Same Survey." *The American Statistician* 37:319-320.

Silver, Nate. 2017. “The Invisible Undecided Voter,” *FiveThirtyEight*. Accessed at: https://fivethirtyeight.com/features/the-invisible-undecided-voter/

Silver, Nate. 2021. “The Death of Polling is Greatly Exaggerated,” *FiveThirtyEight*. Accessed at: https://fivethirtyeight.com/features/the-death-of-polling-is-greatly-exaggerated/

Smith-Schoenwalder, C. (2020). “Trump Attacks 'Suppression Polls' That Show Biden Ahead,” U.S. News & World Report. https://www.usnews.com/news/elections\/articles/2020-10-28/\trump-attacks-suppression-polls-that-show-biden-ahead.

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

Appendix A: Notes on Data Collection

A tremendous thanks goes to Sarah Lentz at the University of Pennsylvania who did an amazing job collecting, organizing, and managing the data collection efforts. Sarah was assisted by University of Pennsylvania undergraduates Ami Ikuenobe and Grayson Peters. Mellissa Meisels and Sara Kirshbaum at Vanderbilt University also provided invaluable assistance in some of the data collection and analysis as did Rezwana Uddin at NBC News. Our work would not be possible without the dedication and assistance of these amazing individuals.

All of the analyses were done using R so that everything could be replicated. Although we focus mainly on polls with a field period ending in the last two weeks prior to Election Day, for completeness, we collected information on every public poll we were able to find.

For some of the analyses in part two we were fortunate to have several organizations share their individual-level data with us. Our work relies heavily on their generosity.

Appendix B: Other Figures and Tables

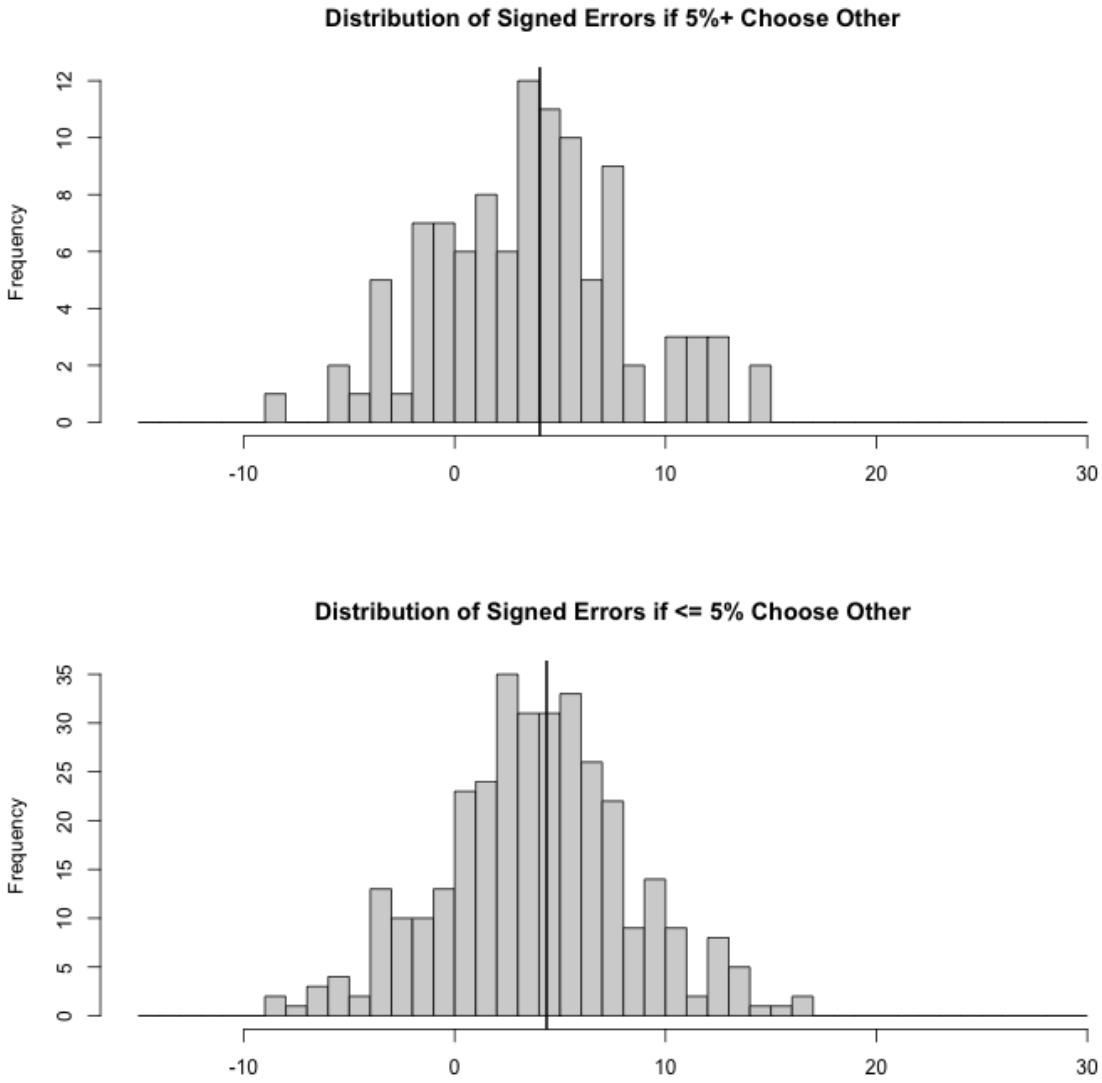


Figure B1. Distribution of Signed Error for Polls with a Larger (>5%) and Smaller (<=5%) Proportion of Respondents Choosing a Response Other Than Biden or Trump. Means are reported by the vertical line.

Type of Poll	Percent Identified Winning Candidate: Strict	Percent Identified Winning Candidate: Poll Margin > MoE	Percent Identified Winning Candidate: Vote Margin > MoE
National Presidential Popular Vote	100 (N=66)	100 (N=66)	100 (N=66)
State-Level Presidential Vote	76.3 (N=438)	90.9 (N=312)	84.5 (N=264)
State-Level Presidential Vote in a State with a Gubernatorial or Senatorial Contest	79.3 (N=261)	95.6 (N=136)	91.7 (120)
Gubernatorial and Senatorial Polls	69.1 (N=181)	84.6 (N=123)	80.2 (96)
Senatorial Polls	65.6 (N=157)	82.2 (N=101)	75.7 (N=74)
Gubernatorial Polls	91.7 (N=24)	95.5 (N=22)	95.7 (N=23)

Table B1. Percent of Polls that Correctly Identified Election Winner from Identified Polls Conducted in the Last Two Weeks by Contest Type.

	Dependent variable:	
	Signed Error	Absolute Error
Trump 2016 Vote	0.26***	0.19***
	(0.09)	(0.07)
Same Day Registration State?	0.48	1.03
	(1.26)	(0.98)
Early Vote as % of 2016 Total	0.01	0.003
	(0.02)	(0.02)
Cost of Voting Rank	-0.03	-0.06*
	(0.05)	(0.04)
Population Density	0.01**	0.01***
	(0.002)	(0.002)
log Median Income	-3.00	-7.99
	(8.71)	(6.75)
% Bachelors	0.15	0.11
	(0.21)	(0.16)
% Black	-0.25***	-0.15***
	(0.06)	(0.05)
% Over 65	0.26	0.06
	(0.29)	(0.22)
log(COVID per capita)	81.94	95.38**
	(49.21)	(38.15)
Constant	15.41	80.56
	(96.99)	(75.19)
Observations	49	49
R^2	0.54	0.56
Adjusted R^2	0.42	0.45
Residual Std. Error (df = 38)	3.03	2.35
F Statistic (df = 10; 38)	4.54***	4.94***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table B2. Predicting Average Signed Error and Average Absolute Error in State Presidential Polls by State. Full set of covariates.

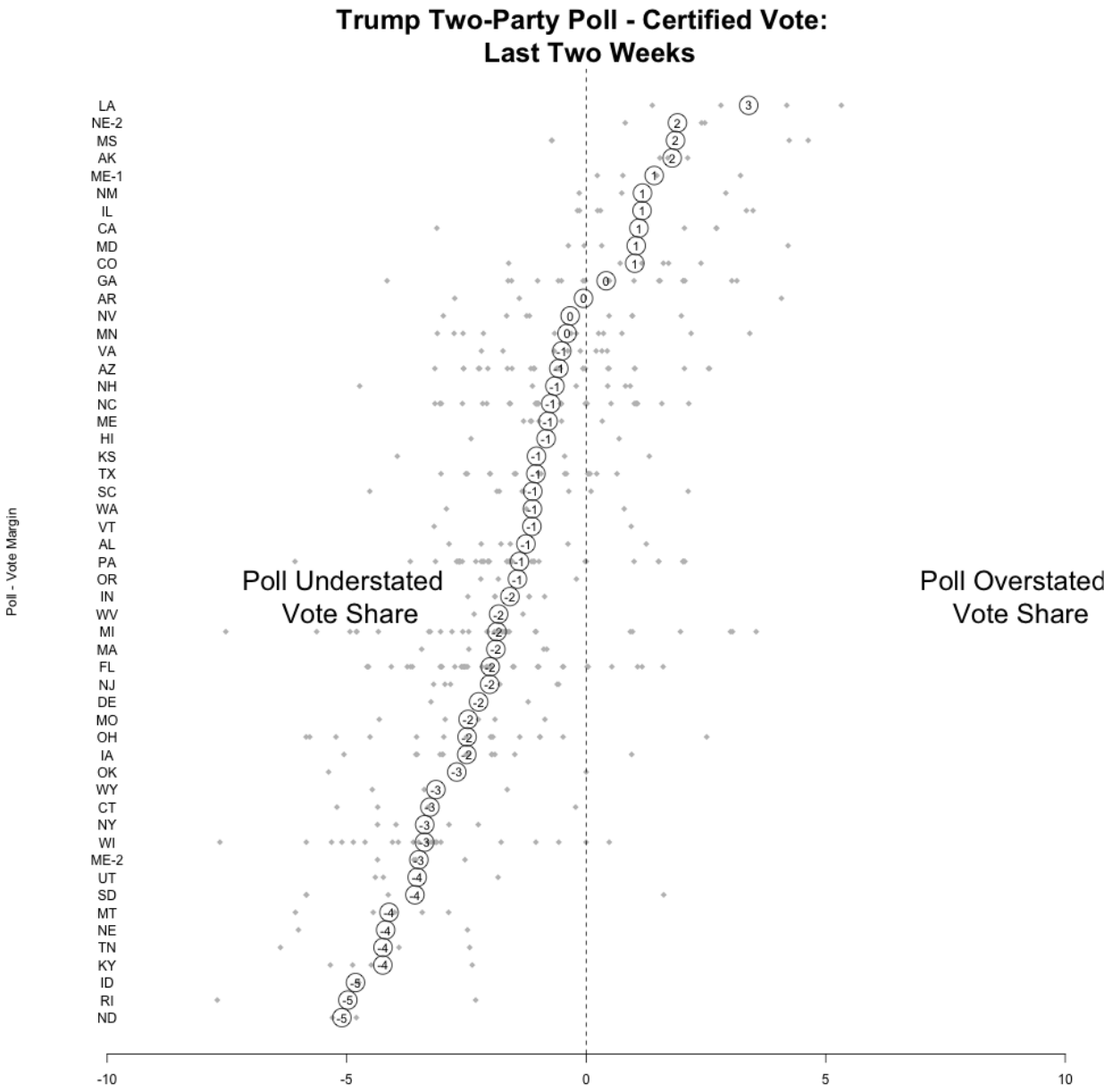


Figure B2-1. Polling Error on Trump by State Using the Two-Party Vote. All polls with a field period ending between 10/21/20 and 11/3/20 are included. Circles denote the average polling error on Trump's vote share for each state (reported within each circle). This analysis calculates the candidate percentages taking out respondents who gave a response other than Trump or Biden. Trump error is computed as Trump's percentage in the poll minus Trump's percentage in the certified vote. Results are sorted by the average Trump error in the state. Individual polls are plotted in grey points.

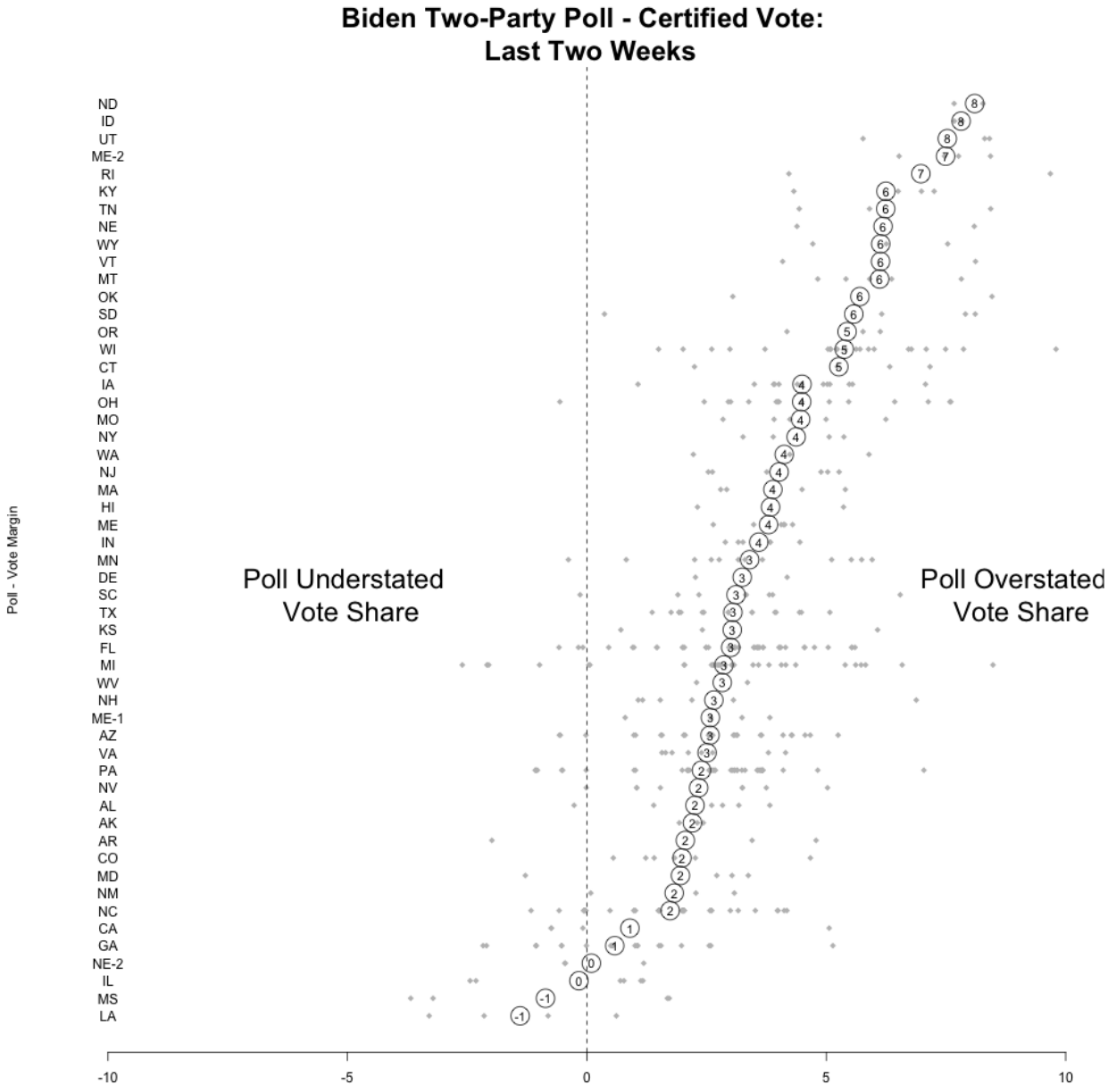


Figure B2-2. Polling Error on Biden by State Using the Two-Party Vote. All polls with a field period ending between 10/21/20 and 11/3/20 are included. Circles denote the average polling error on Biden's vote share for each state (reported within each circle). This analysis calculates the candidate percentages taking out respondents who gave a response other than Trump or Biden. Biden error is computed as Biden's percentage in the poll minus Biden's percentage in the certified vote. Results are sorted by the average Biden error in the state. Individual polls are plotted in grey points.

Appendix C. Effects of Excluding Undecideds, Focusing on Competitive States, and Using "Mixed" Modes

The conclusion about polling modes is not changed by the following: excluding undecideds (i.e., those respondents who don't select a candidate in the vote question) to remove question-wording differences, focusing only on polls conducted in competitive states, and using mixed-mode surveys versus single methods of interviewing.

One way the mode of interviewing matters is in how easily respondents indicate uncertainty about their vote. For example, unlike phone polls that have a human interviewer to record a "don't know" response, online polls may record a "don't know" response if the respondent skips the question. While the overall percentage of respondents who give an answer other than a major-party candidate was relatively low in 2020 (Table 2), we replicated the analysis using the two-candidate vote share to remove respondents who did not provide a candidate preference.[52] Table C1 shows that this did not affect the conclusions of Figure 8.

	Signed Error	Absolute Error	Pct Correct Winner (Simple)	Number of Polls
Live Phone - RDD	6.0	6.2	67	27
Live Phone - RBS	5.4	5.4	70	31
Online	4.7	5.3	82	262
IVR/Online	3.6	4.6	71	42
Phone/Online	2.8	4.9	56	32
Other Mixed/Misc.	3.5	5.6	68	44

Table C1. Comparing Mode Effects for State-Level Presidential Polls Conducted in Last Two Weeks after Removing Don't Knows.[53] Individual poll results are renormalized so that support for Biden and Trump is a percentage relative to the percentage of respondents supporting either Biden or Trump in the poll before computing the polling error and taking the average by mode. All polls with a field period ending between 10/21/20 and 11/3/20 are included. Only modes with at least 10 polls are included.

Previously Figure 2 showed that the phone polls were more likely to be undertaken in competitive states. To ensure that mode effects are not driven by differences at the state level, Table C2 shows that similar conclusions emerge when only analyzing competitive states (i.e., states where the final margin was five percentage points or less).

[52] Thus, if 49% supported Biden and 43% supported Trump, marginals could be renormalized to remove the 8% who didn't give a response as follows: 49% / 92% = 53.3% for Biden and 43% / 92% = 46.7% for Trump.

[53] The renormalization can also produce misleading results. For example, if 50% of the electorate declines to state a preference and the remaining 50% is split equally between the candidates, how should we interpret the results? The conventional interpretation would suggest that 25% support each candidate. The normalization used to create Table C1 would suggest 50% support for each candidate. The normalization is useful for comparing across modes with varying levels of "don't knows" due to the interviewing mode, but it seems hard to justify it as a general practice.

	Signed Error	Absolute Error	Pct Correct Winner (Naïve)	Number of Polls
Live Phone - RDD	5.3	5.6	67	21
Live Phone - RBS	5.0	5.0	74	23
Online	4.6	4.8	71	120
IVR/Online	3.5	4.3	77	26
Phone/Online	1.2	3.9	52	25
Other Mixed/Misc.	4.0	5.1	63	24

Table C2. Comparing Mode Effects for State-Level Presidential Polls Conducted in Last Two Weeks in States with 2016 Vote Margins of Five Points or Less. Competitive states are those where the margin between Trump and Clinton in 2016 was less than or equal to five percentage points. All polls with a field period ending between 10/21/20 and 11/3/20 are included. Only modes with at least 10 polls are included.

Pollsters were more likely to use multiple modes of interviewing in 2020 compared to prior elections. The performance, on average, of surveys via telephone (either RBS or RDD) and online surveys were compared to the performance of polls that used multiple interviewing modes. We are hesitant to make strong claims since it is impossible to identify how important the mode of interview is for polling performance, considering all of the other potential confounding differences. Hence, this analysis is presented for descriptive purposes only.

To perform this evaluation, all of the multi-mode polls were collapsed into a single "mixed" category then compared to polls conducted via phone (both RDD and RBS) and online using state-level presidential polls according to the metrics used in the report.

	Number of Polls	Average Signed Error	Average Absolute Error	Percent Correct Winner (Naïve)
"Mixed" mode	74	3.0	4.4	65
Online only	262	4.5	5.2	82
Phone only (RBS or RDD)	58	5.4	5.5	69
All	438	4.3	5.1	76

Table C3. Comparing Mixed and Pure Interview Modes for State-Level Presidential Polls. All polls with a field period ending between 10/21/20 and 11/3/20 for which the interview mode could be determined are included.

Table C3 confirms the pattern that was documented earlier in this report. Polls using mixed interviewing methods have lower signed error than polls conducted exclusively online or exclusively using human telephone interviewer. Furthermore, the absolute error across the modes is even more similar, and the "mixed" polls performed the worst in terms of naïve predictions (which ignore the margin of error). *Mixed* polls have lower signed error and lower naïve classification rate. Nine of the 74 mixed polls (i.e., 12% of the polls) estimated Trump would win; in contrast, only three *Online* polls (1%) and one *Phone* poll (2%) estimated that Trump would win.

These differences are difficult to assess. They could reflect more types being reached by mixed mode polls than single mode polls, the effects of the statistical adjustments and assumptions, or differences in the types of races.

Appendix D: Outcome Reweighting

To perform the reweighting to the 2020 vote outcome, the composition county-level votes was computed using the TargetSmart voter file, and the county level vote for each candidate also was collected from state sources. The past vote of the counties within each state was used to classify all counties into five strata as follows: Very Republican, Lean Republican, Toss-Up, Lean Democratic, Strong Democratic.

Thus, for each state, the reweighting was performed as follows.

1) Use the original weights to calculate the proportion of respondents in each category of potential interest.
2) Calculate the proportion of respondents who indicated voting via absentee, early in-person or in-person on Election Day, using the original poll weights in each of the five strata.
3) Using the voter file, calculate the proportion of overall votes cast, using each voting method in each stratum.
4) Generate a new weight using the inverse of the ratio of steps 2 and 3 and create a new weight by multiplying the original weight by this new adjustment.
5) Using this new weight, calculate the proportion of respondents who support Biden and Trump in each stratum. Calculate the proportion of votes cast for Biden and Trump in each stratum and create the adjustment using the inverse ratio of the proportions.
6) Create the final weight by multiplying the method-adjusted weight by the outcome-adjusted ratio and use this weight to calculate the proportion of respondents in each category of potential interest.
7) Calculate the difference in the category proportions in step 6 relative to the category proportions in step 1.

Appendix E: Location of Republican Understatement in Outcome Reweighted Polls

> *Reweighting the available individual-level data to match the actual vote allows us to compare how the original proportion of Trump support in a county-based party stratum compares to the proportion estimated after reweighting the data to match the true proportion of vote (and method of voting) in that stratum. This approach can help identify whether the understatement of Trump support was more likely to occur in Democrat-leaning counties, Republican-leaning counties, or counties that were closely contested.*

Results weighted to county-level partisan strata allow for the adjustments to errors to be examined in terms respondents living in more Republican or more Democratic counties. Although inconclusive because Biden supporters live in Republican counties and Trump supporters live in Democratic counties, such analysis helps describe the nature of the polling error in 2020.

To do this analysis, originally weighted electorate in each of the party strata in each state is compared to the electorate implied after reweighting the data to match the outcome. The proportion of votes cast in each stratum in the outcome reweighting can be compared with the original survey weights. Whether the original survey weights overrepresented or underrepresented voters in particular counties can be estimated from this comparison. For example, if the originally weighted data understated the proportion of voters in the most-Republican counties, because of nonresponse or weighting adjustments, then proportion of respondents living in those counties would appear to increase after the reweighting.

For the proportion of Trump votes cast in each stratum, Figure E1 plots by state the difference between the original poll weights and the outcome-adjusted reweighting.[54] To explain Figure E1, consider the case of Wisconsin. The results suggest that the percentage of respondents living in High Republican counties is roughly unchanged after the reweighting (in fact it slightly decreases). Hence, the original weights had roughly the correct percentage of voters from those counties. In contrast, the difference is around three percentage points for Lean Republican counties, meaning that the originally weighted poll had too few respondents from counties that leaned Republican relative to the number required to make the poll results match the election results.

The relationship varies somewhat across states. The largest understatement often occurs in toss-up or Republican-leaning counties. In particular, the states with the largest understatement in Republican-leaning counties include WI, MT, KS, IA, and NC. The states with the largest understatement in Toss-up counties include MI, PA, and FL. It is generally not the case that the largest understatement occurs in the most Republican or most Democratic counties.

[54] Because the strata sizes vary across states, the differences sum to one only after weighting by strata size.

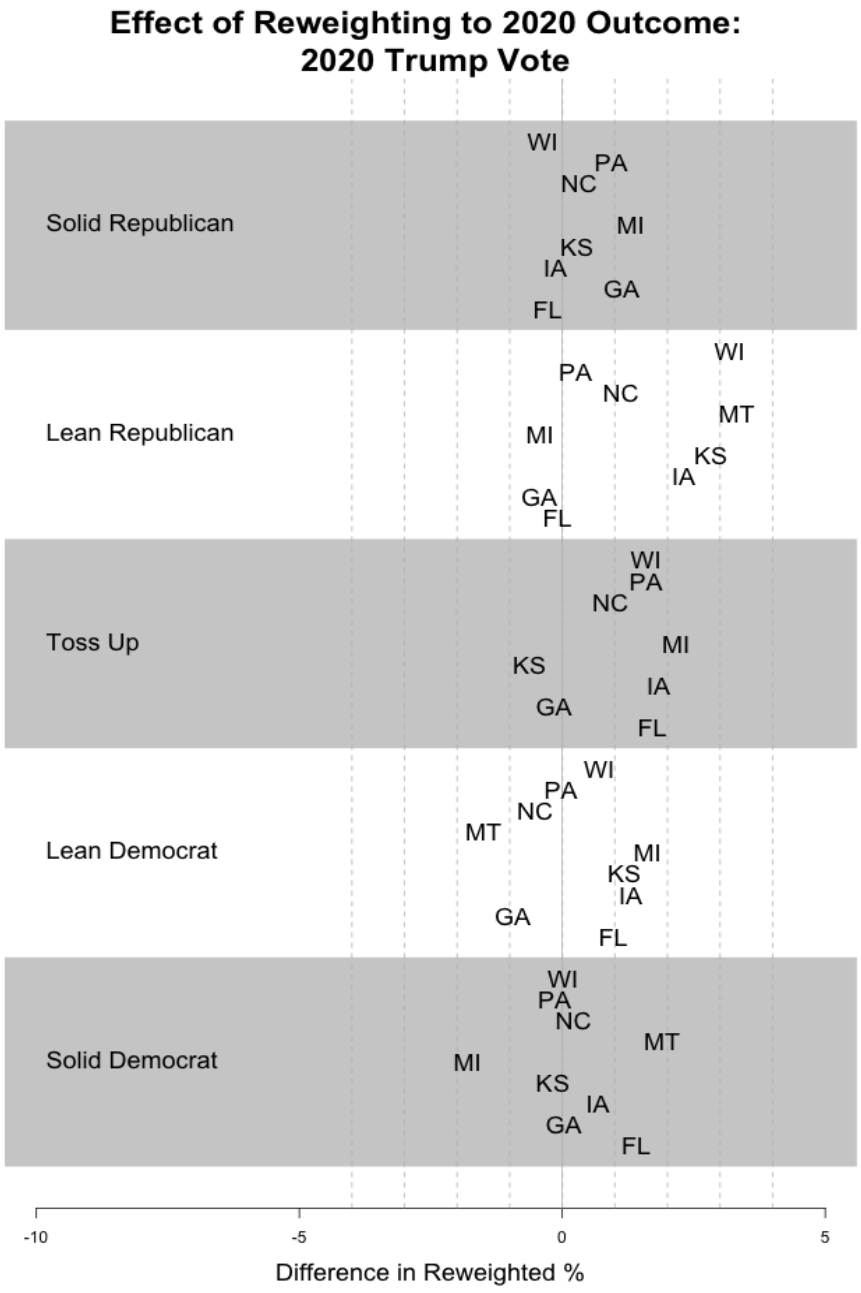
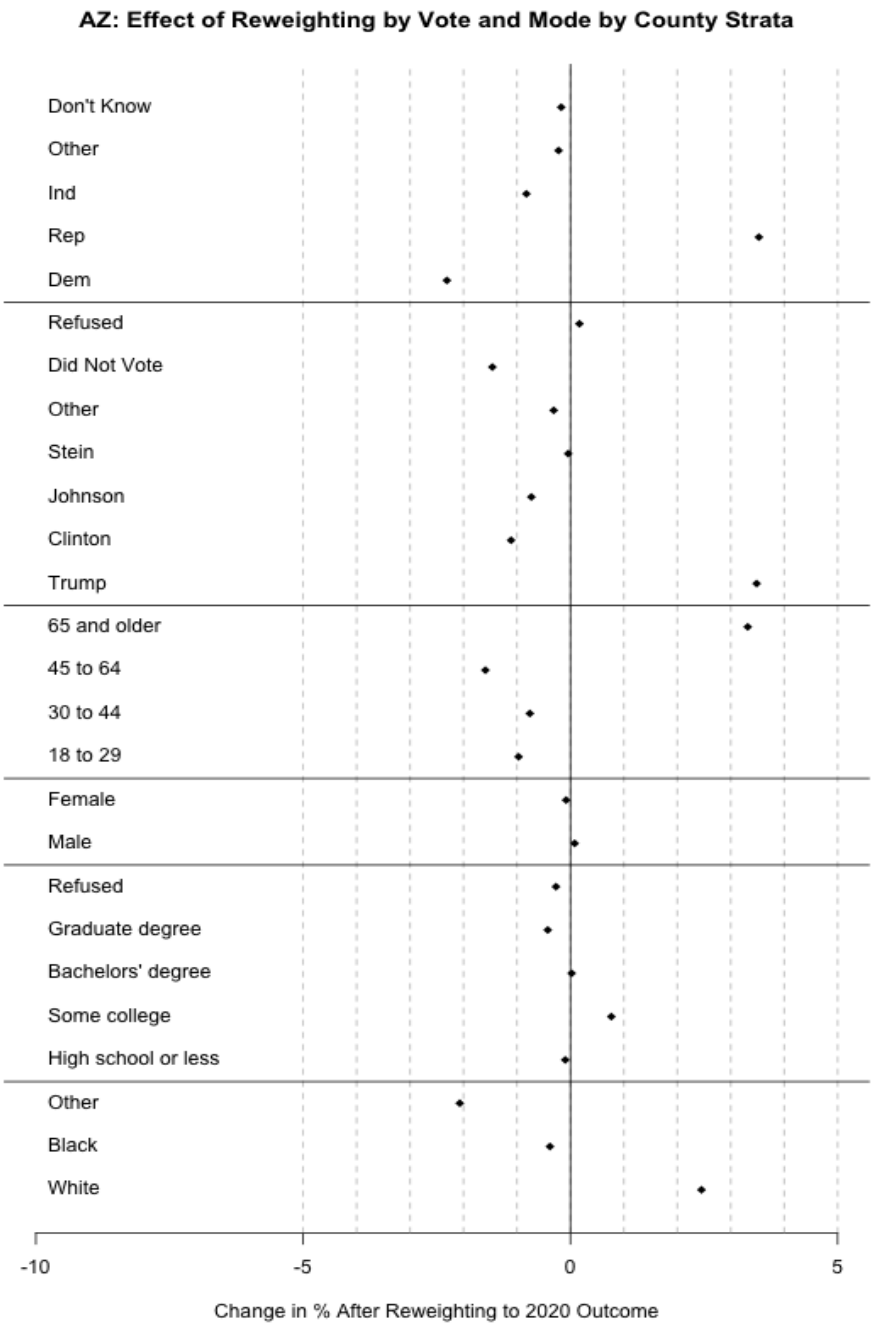
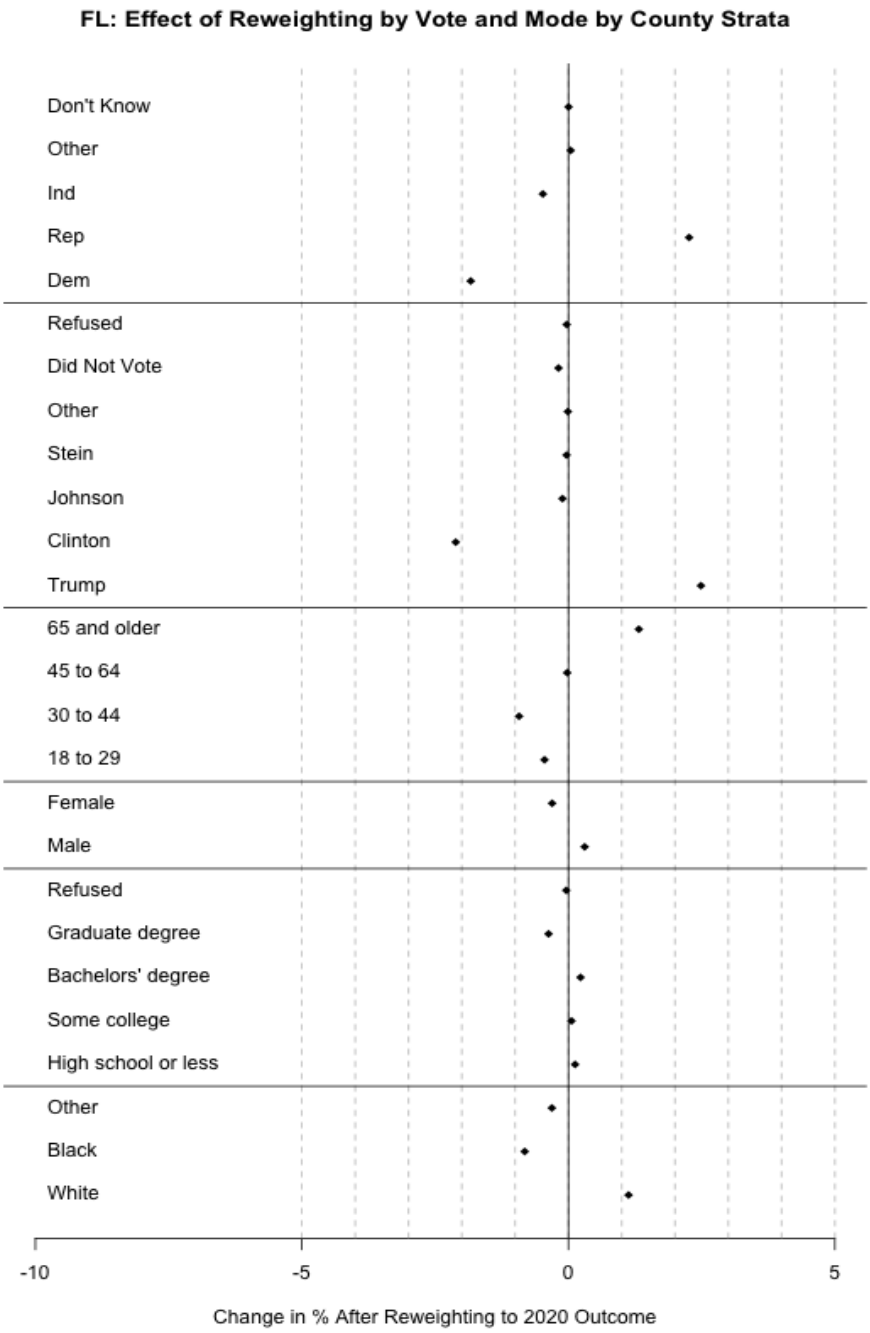
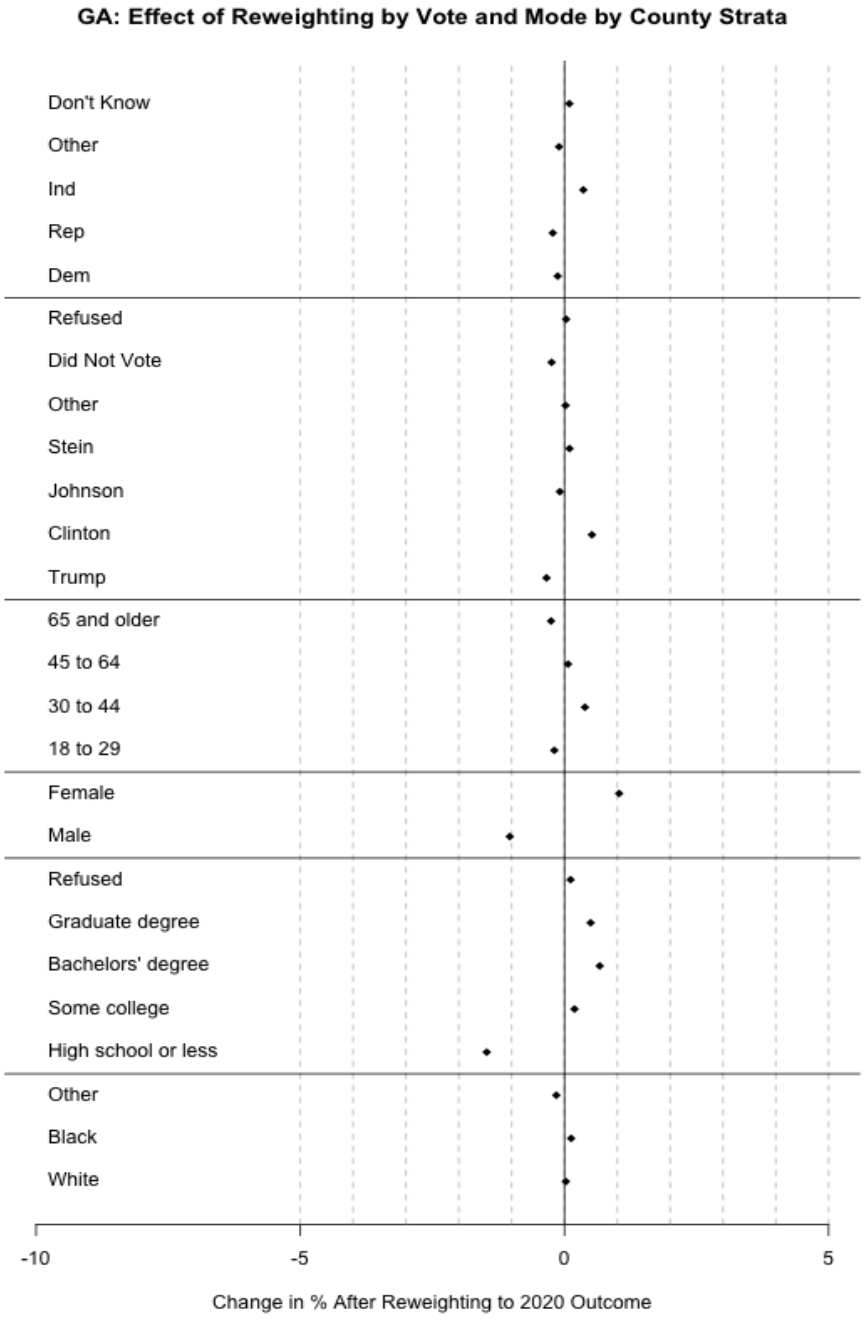


Figure E1. The Change in Electoral Composition for County Party Strata. Plotted data points represent the differences between the original weights and outcome-adjusted weights as applied to estimate the proportion of the electorate in each party stratum at the county level. Positive values indicate that the outcome-adjusted weights estimate more Trump votes than the original weighting. The polls analyzed are RBS telephone polls.

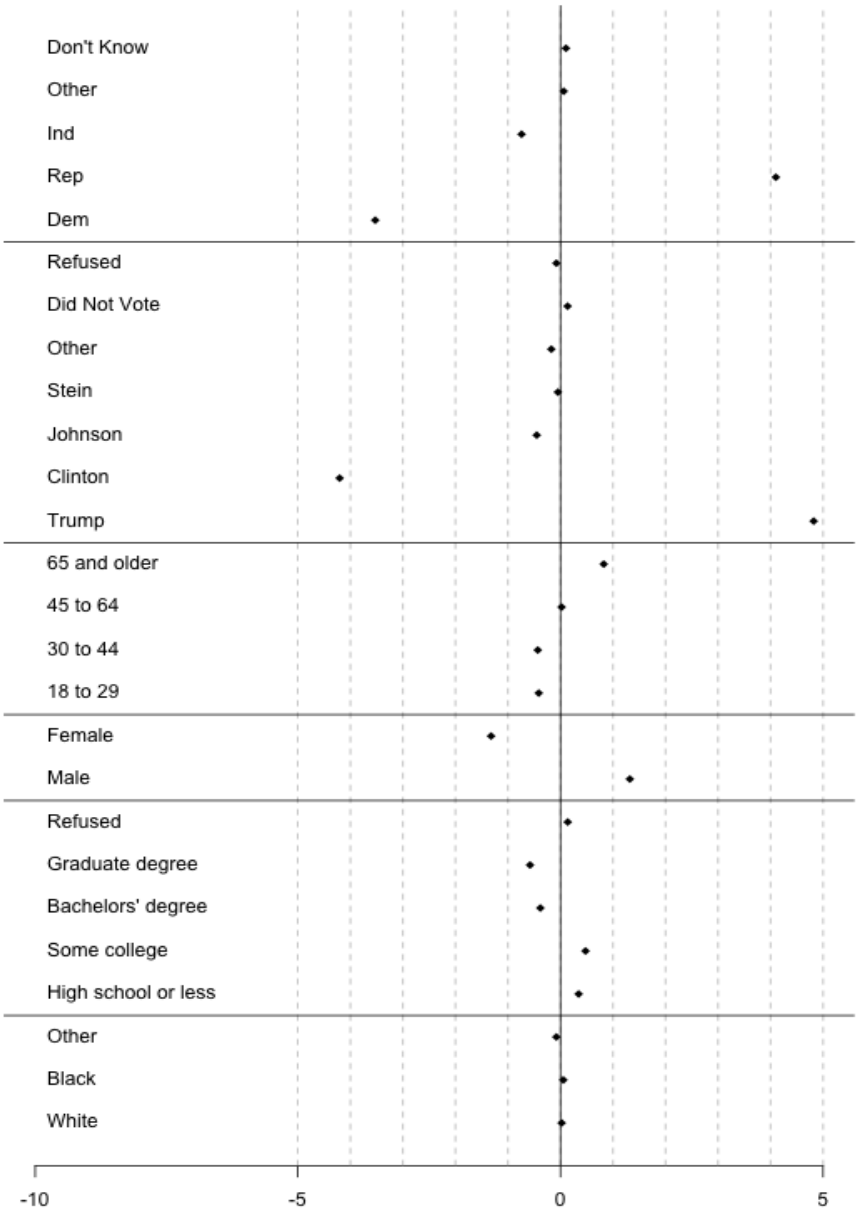
Appendix F:
Results of Reweighting Individual State RBS Polls

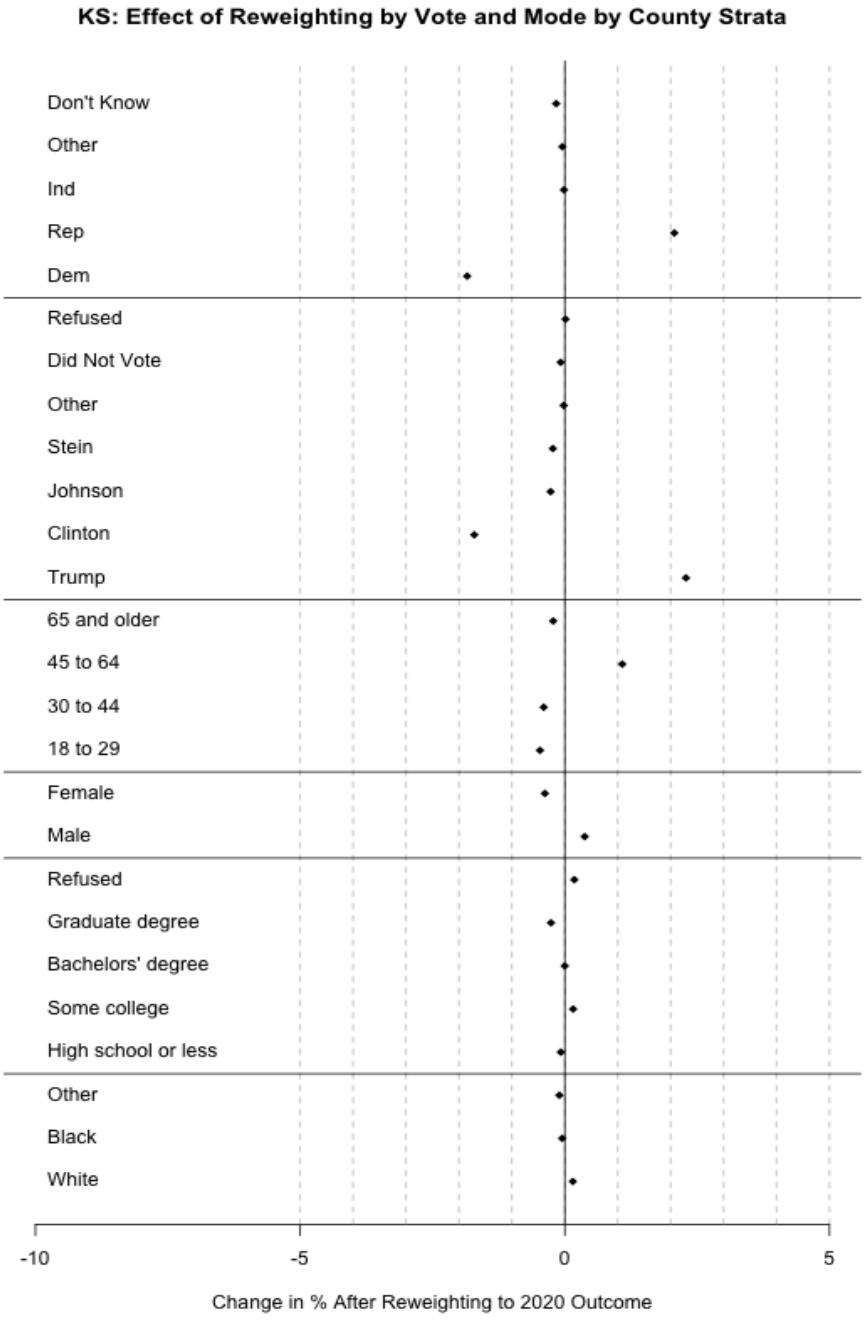


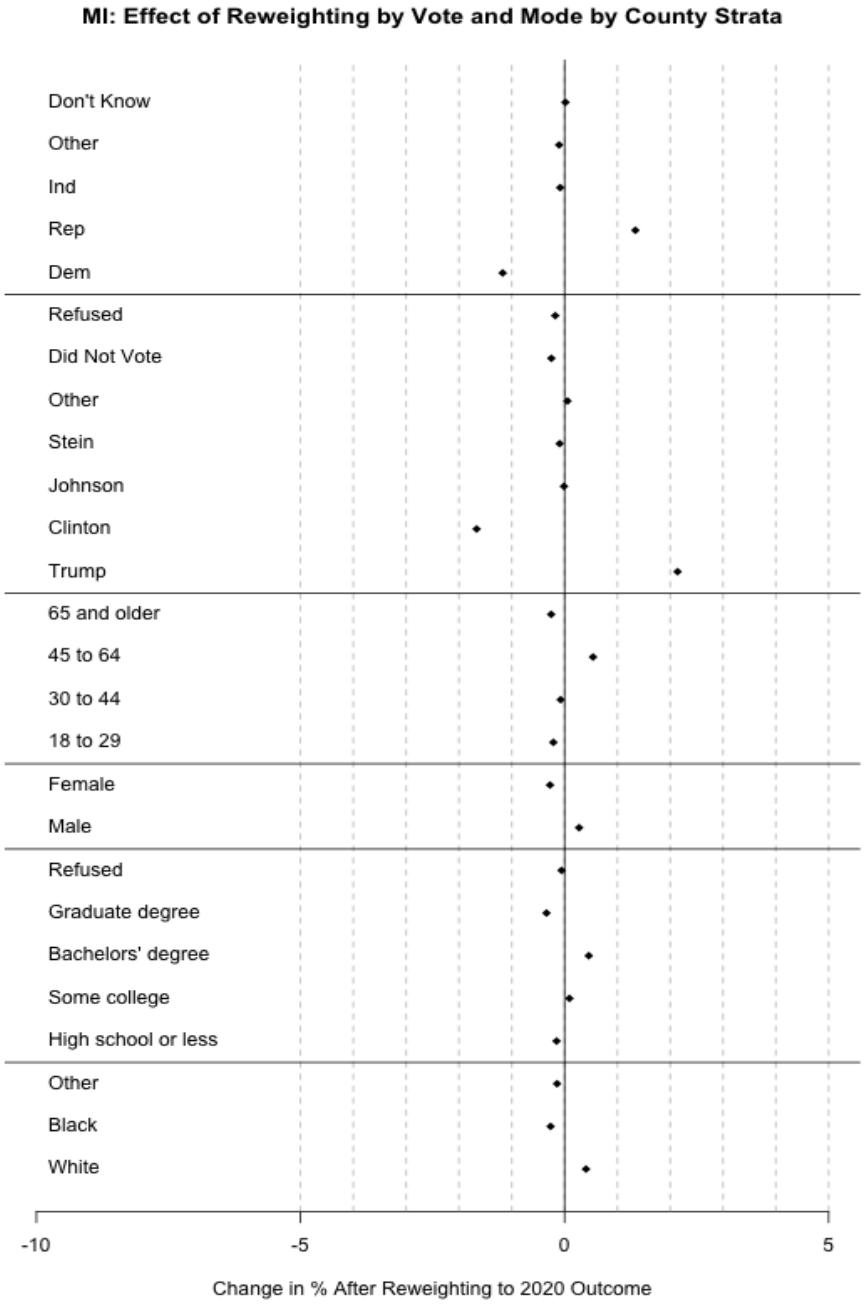

FL: Effect of Reweighting by Vote and Mode by County Strata
Don't Know
Other
Ind
Rep
Dem
Refused
Did Not Vote
Other
Stein
Johnson
Clinton
Trump
65 and older
45 to 64
30 to 44
18 to 29
Female
Male
Refused
Graduate degree
Bachelors' degree
Some college
High school or less
Other
Black
White
-10
-5
0
5
Change in % After Reweighting to 2020 Outcome

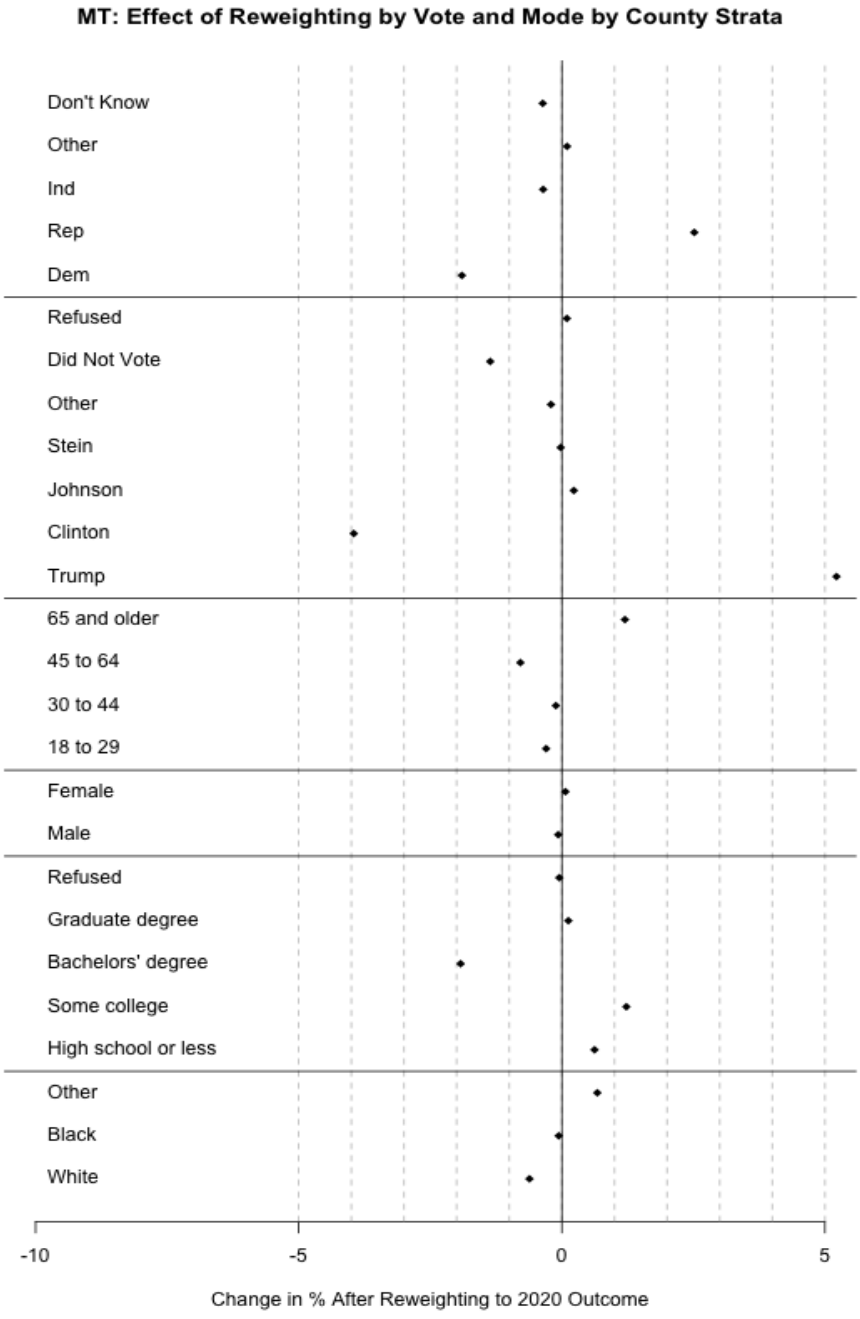

GA: Effect of Reweighting by Vote and Mode by County Strata
Don't Know
Other
Ind
Rep
Dem
Refused
Did Not Vote
Other
Stein
Johnson
Clinton
Trump
65 and older
45 to 64
30 to 44
18 to 29
Female
Male
Refused
Graduate degree
Bachelors' degree
Some college
High school or less
Other
Black
White
-10
-5
0
5
Change in % After Reweighting to 2020 Outcome

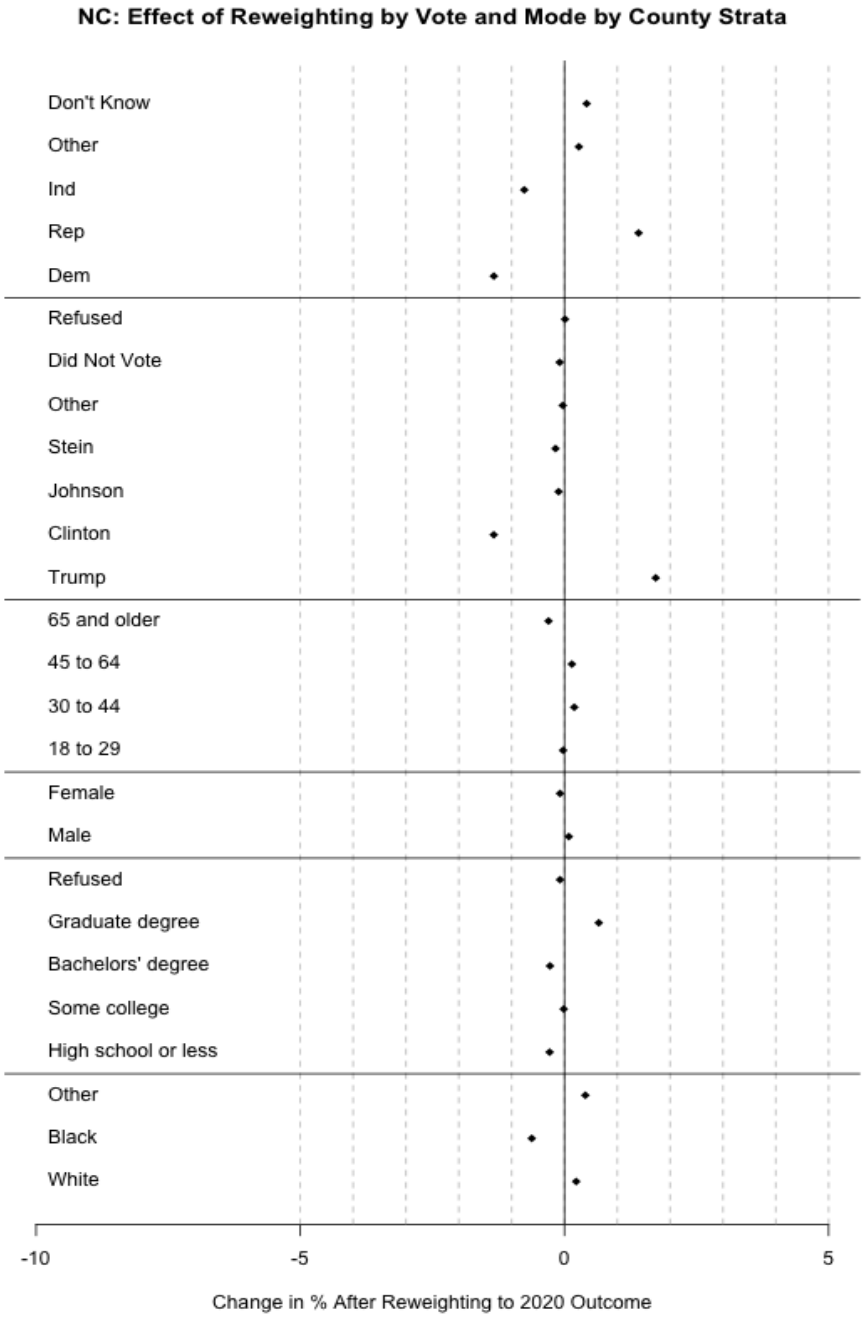
IA: Effect of Reweighting by Vote and Mode by County Strata

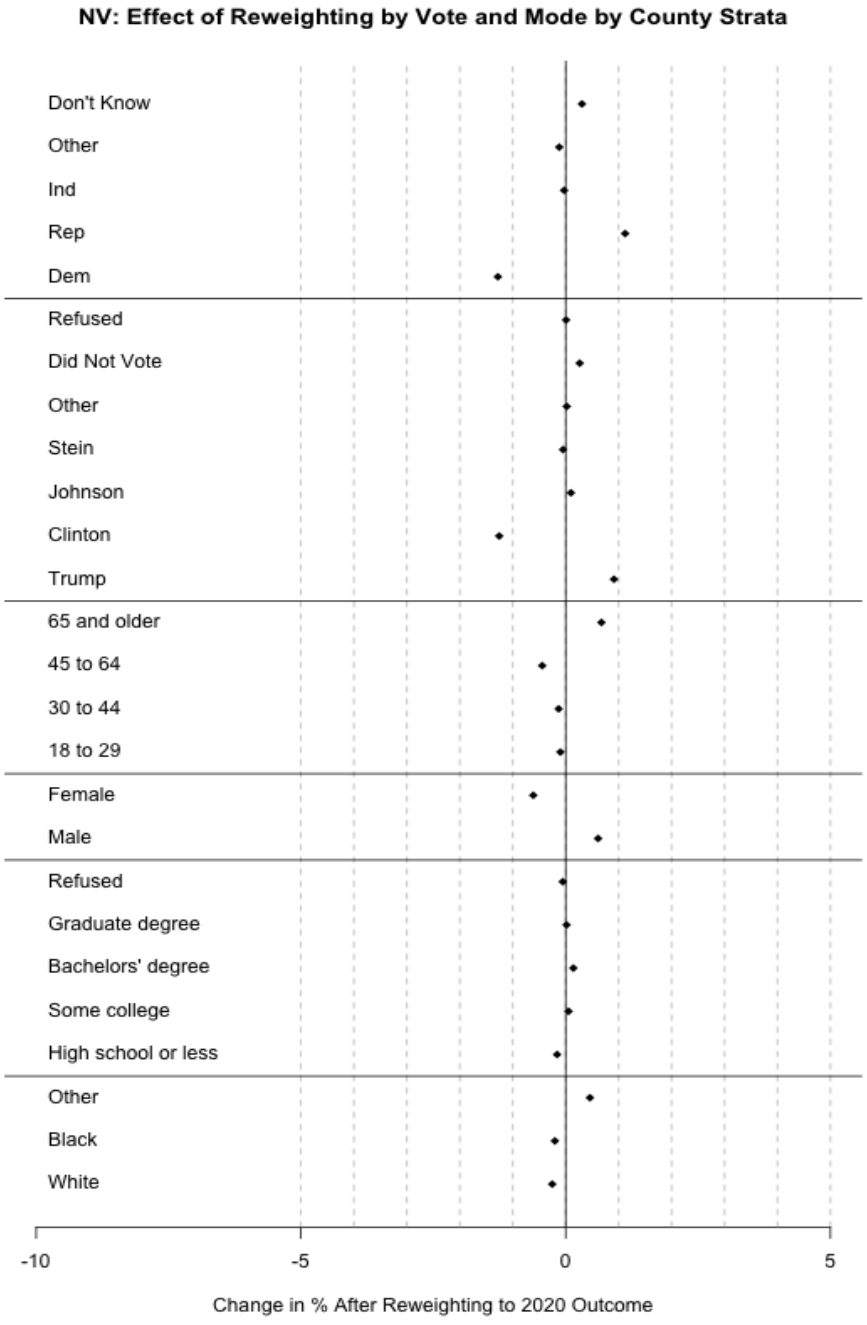


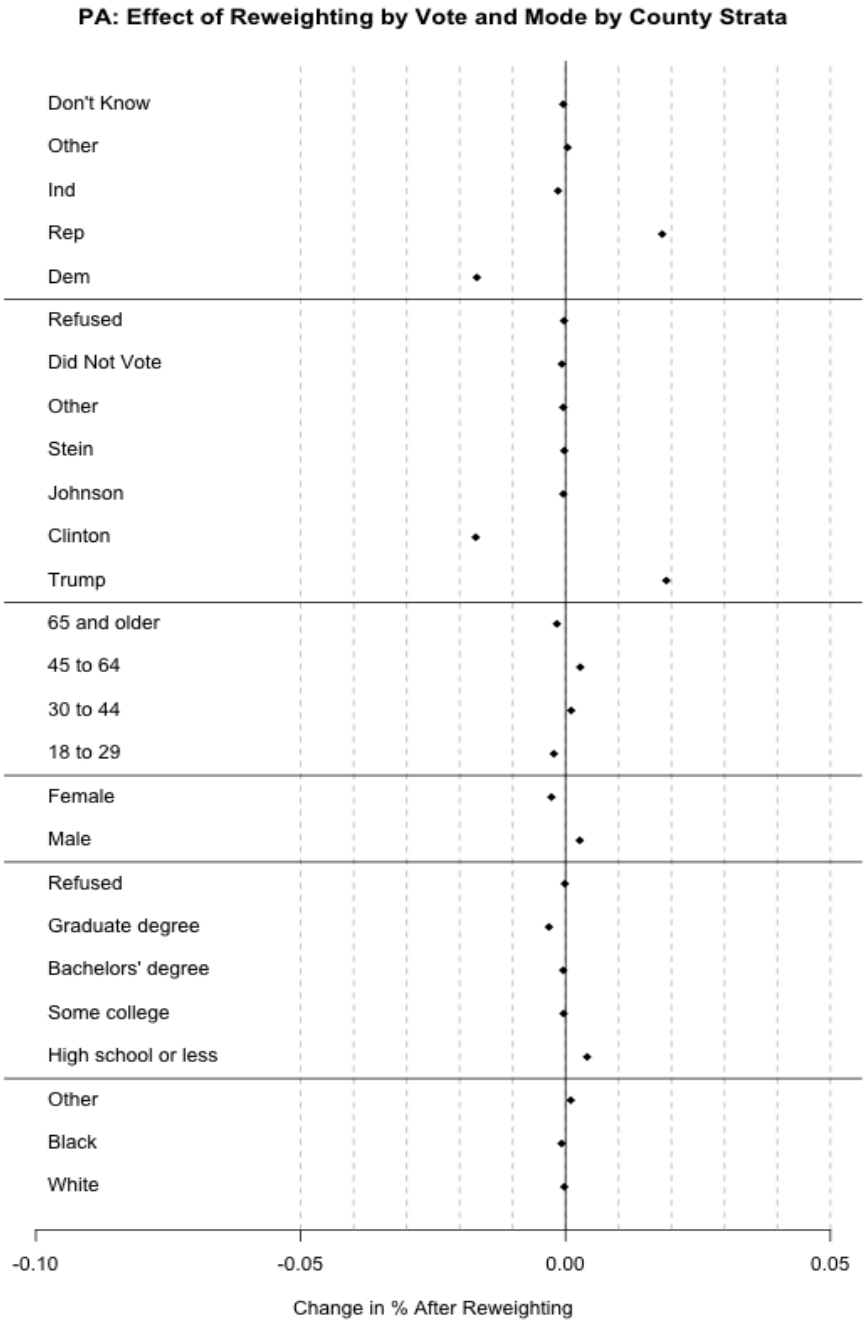

KS: Effect of Reweighting by Vote and Mode by County Strata
Don't Know
Other
Ind
Rep
Dem
Refused
Did Not Vote
Other
Stein
Johnson
Clinton
Trump
65 and older
45 to 64
30 to 44
18 to 29
Female
Male
Refused
Graduate degree
Bachelors' degree
Some college
High school or less
Other
Black
White
-10
-5
0
5
Change in % After Reweighting to 2020 Outcome

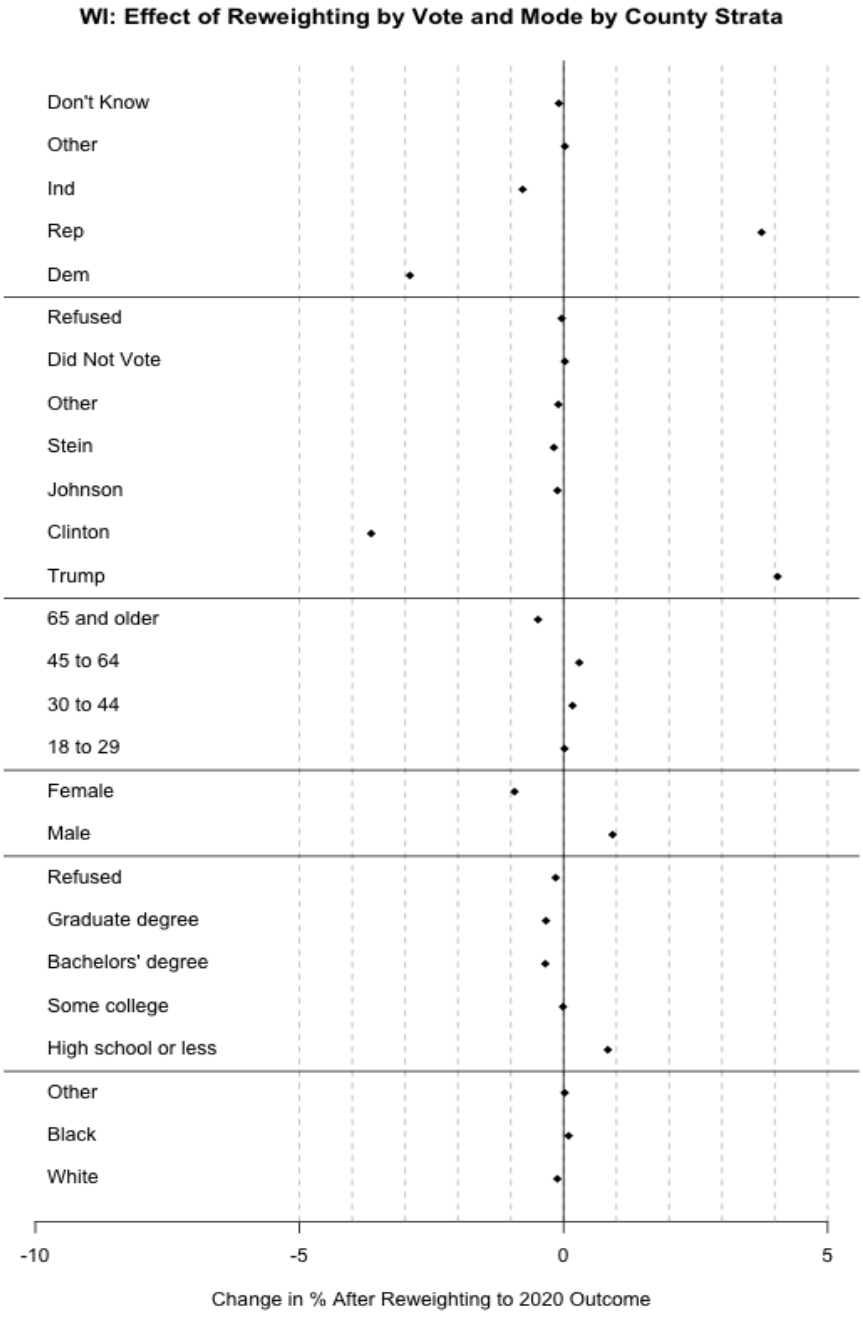

MI: Effect of Reweighting by Vote and Mode by County Strata
Don't Know
Other
Ind
Rep
Dem
Refused
Did Not Vote
Other
Stein
Johnson
Clinton
Trump
65 and older
45 to 64
30 to 44
18 to 29
Female
Male
Refused
Graduate degree
Bachelors' degree
Some college
High school or less
Other
Black
White
-10
-5
0
5
Change in % After Reweighting to 2020 Outcome


MT: Effect of Reweighting by Vote and Mode by County Strata
Don't Know
Other
Ind
Rep
Dem
Refused
Did Not Vote
Other
Stein
Johnson
Clinton
Trump
65 and older
45 to 64
30 to 44
18 to 29
Female
Male
Refused
Graduate degree
Bachelors' degree
Some college
High school or less
Other
Black
White
-10
-5
0
5
Change in % After Reweighting to 2020 Outcome

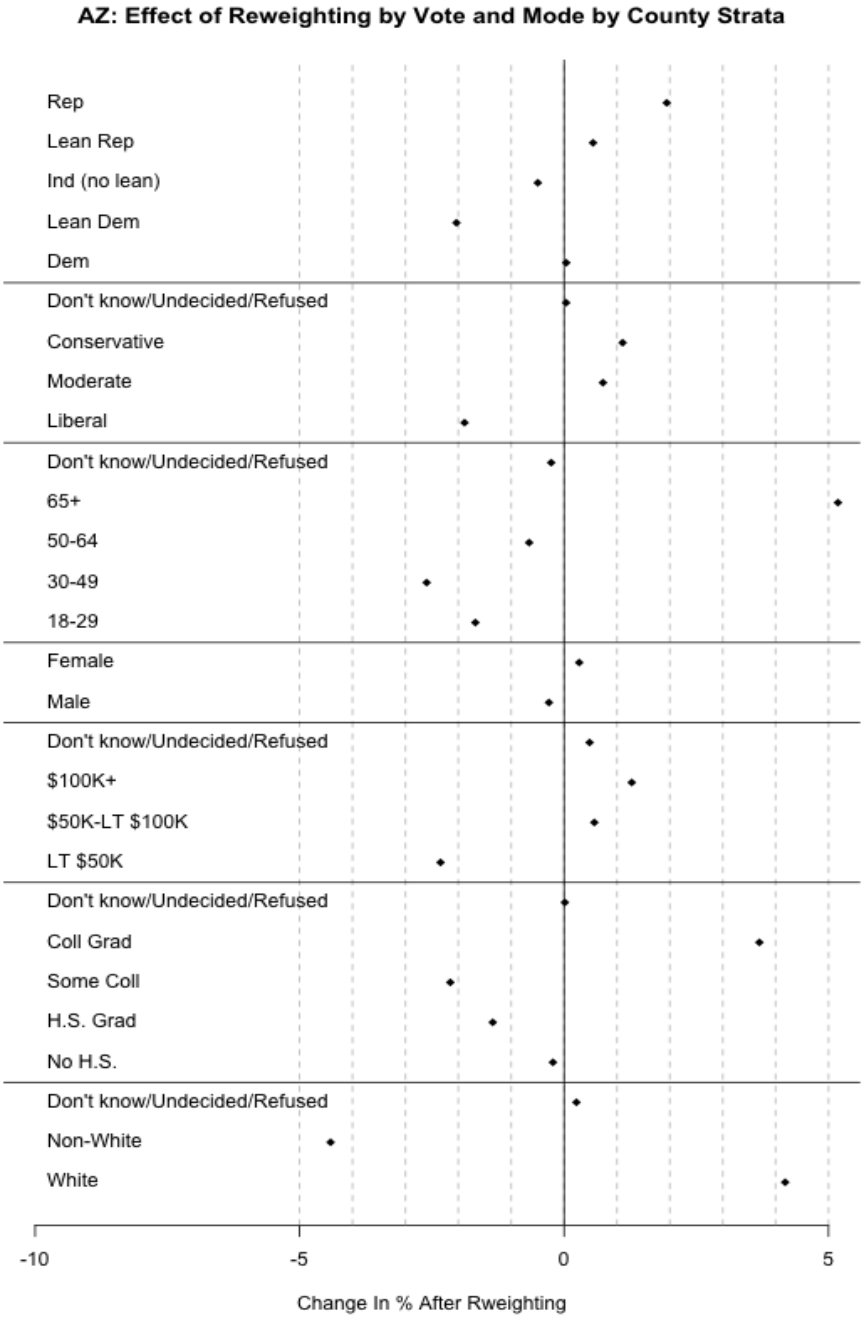

NC: Effect of Reweighting by Vote and Mode by County Strata
Don't Know
Other
Ind
Rep
Dem
Refused
Did Not Vote
Other
Stein
Johnson
Clinton
Trump
65 and older
45 to 64
30 to 44
18 to 29
Female
Male
Refused
Graduate degree
Bachelors' degree
Some college
High school or less
Other
Black
White
-10
-5
0
5
Change in % After Reweighting to 2020 Outcome

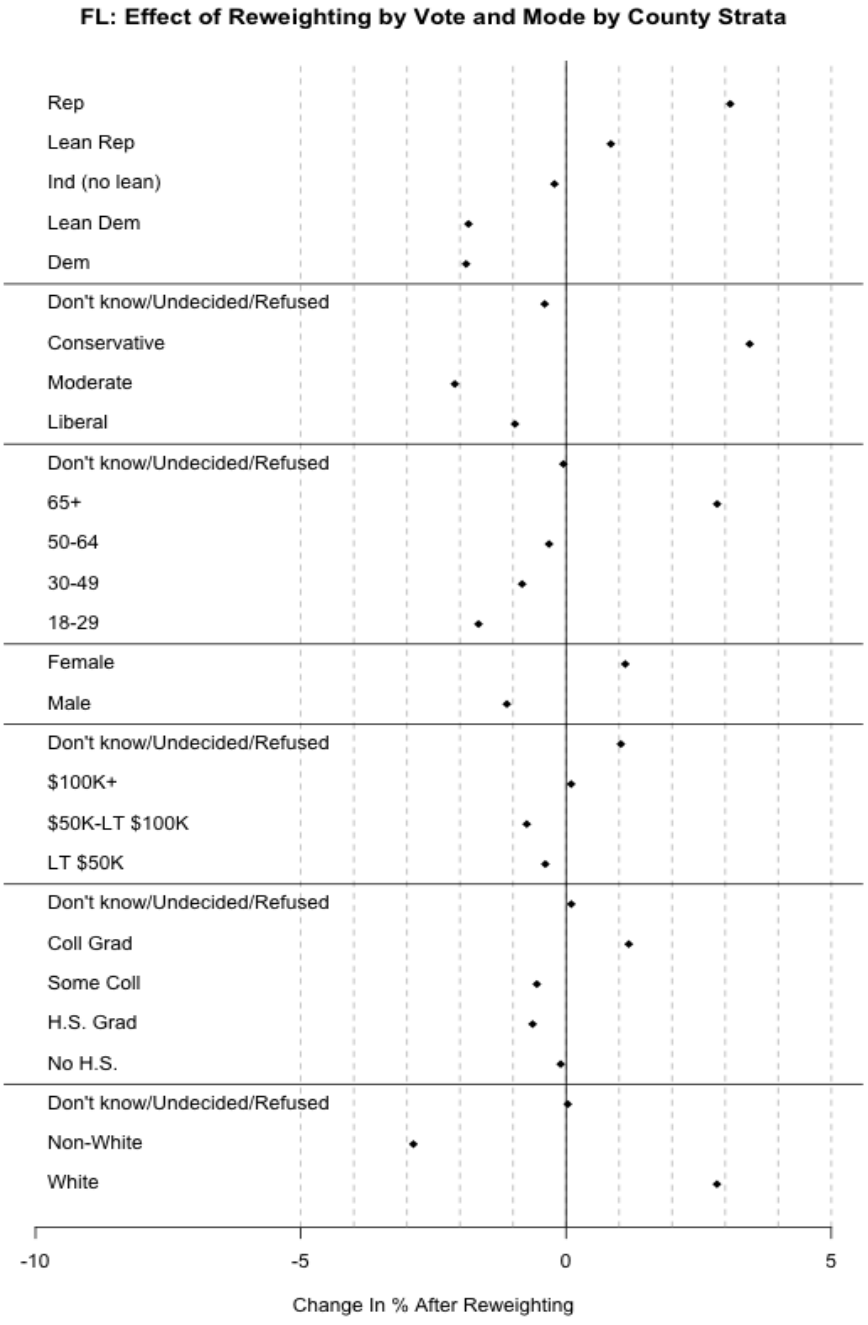

NV: Effect of Reweighting by Vote and Mode by County Strata
Don't Know
Other
Ind
Rep
Dem
Refused
Did Not Vote
Other
Stein
Johnson
Clinton
Trump
65 and older
45 to 64
30 to 44
18 to 29
Female
Male
Refused
Graduate degree
Bachelors' degree
Some college
High school or less
Other
Black
White
-10
-5
0
5
Change in % After Reweighting to 2020 Outcome

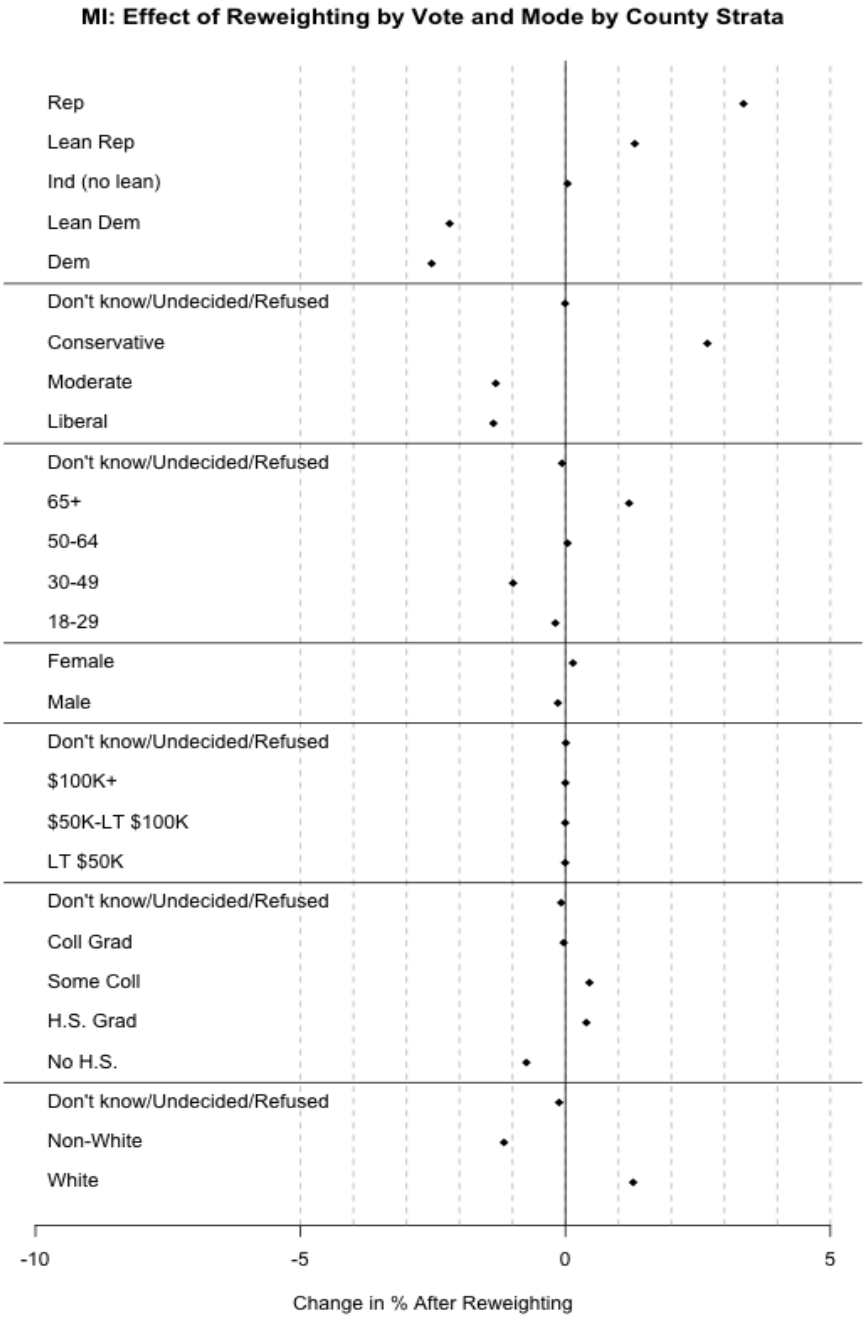

PA: Effect of Reweighting by Vote and Mode by County Strata
Don't Know
Other
Ind
Rep
Dem
Refused
Did Not Vote
Other
Stein
Johnson
Clinton
Trump
65 and older
45 to 64
30 to 44
18 to 29
Female
Male
Refused
Graduate degree
Bachelors' degree
Some college
High school or less
Other
Black
White
-0.10
-0.05
0.00
0.05
Change in % After Reweighting

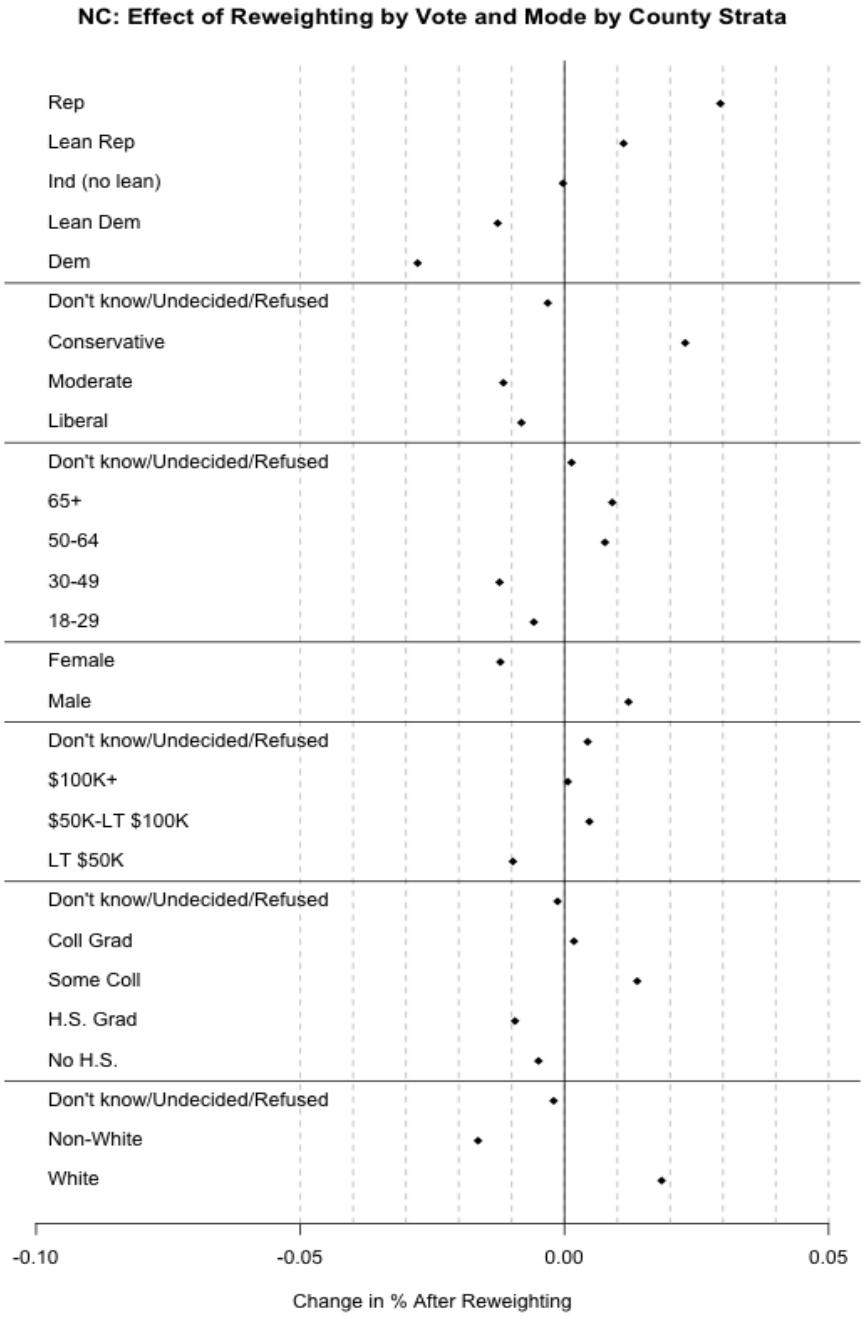

WI: Effect of Reweighting by Vote and Mode by County Strata
Don't Know
Other
Ind
Rep
Dem
Refused
Did Not Vote
Other
Stein
Johnson
Clinton
Trump
65 and older
45 to 64
30 to 44
18 to 29
Female
Male
Refused
Graduate degree
Bachelors' degree
Some college
High school or less
Other
Black
White
-10
-5
0
5
Change in % After Reweighting to 2020 Outcome

Appendix G:
Results of Reweighting Individual State RDD Polls

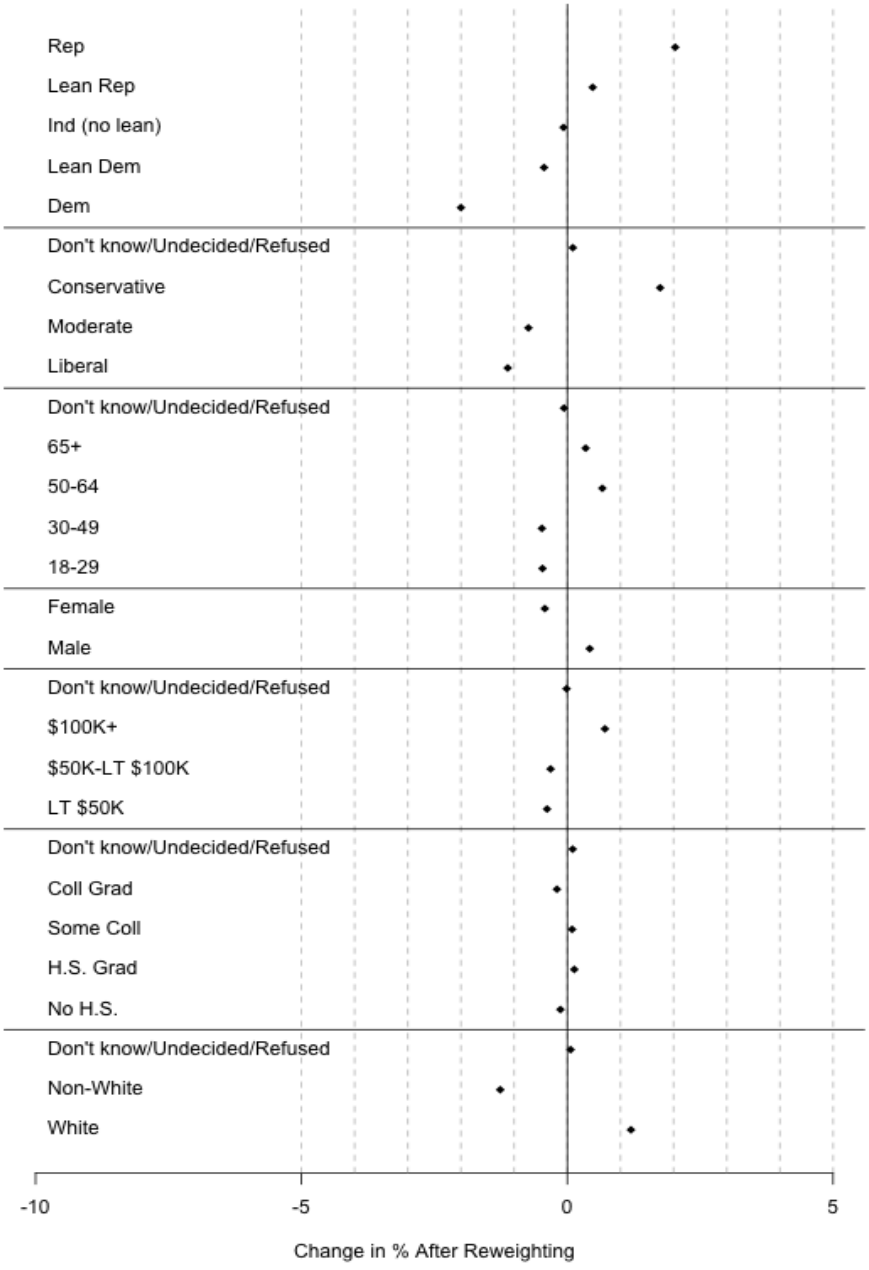


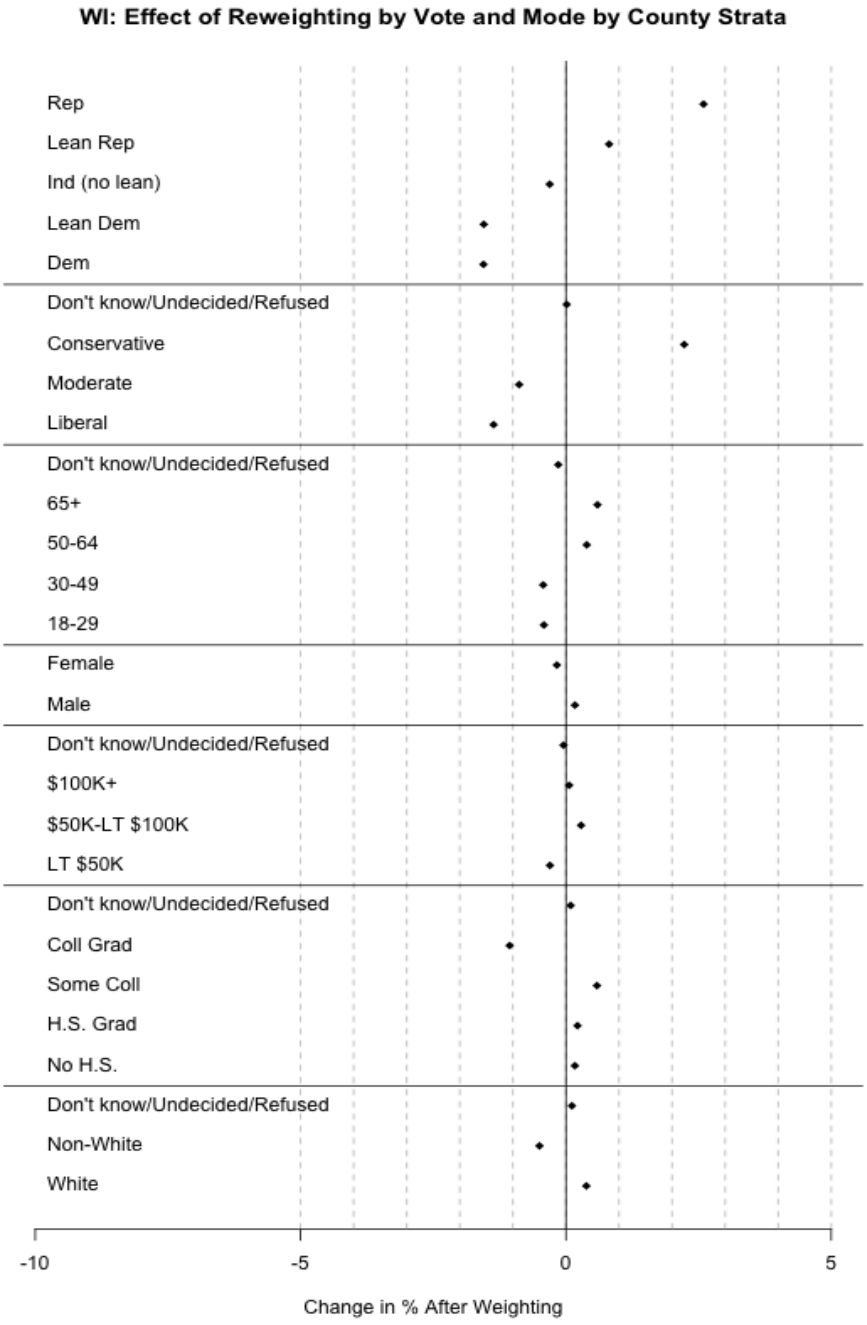

FL: Effect of Reweighting by Vote and Mode by County Strata
Rep
Lean Rep
Ind (no lean)
Lean Dem
Dem
Don't know/Undecided/Refused
Conservative
Moderate
Liberal
Don't know/Undecided/Refused
65+
50-64
30-49
18-29
Female
Male
Don't know/Undecided/Refused
$100K+
$50K-LT $100K
LT $50K
Don't know/Undecided/Refused
Coll Grad
Some Coll
H.S. Grad
No H.S.
Don't know/Undecided/Refused
Non-White
White
-10
-5
0
5
Change In % After Reweighting


MI: Effect of Reweighting by Vote and Mode by County Strata
Rep
Lean Rep
Ind (no lean)
Lean Dem
Dem
Don't know/Undecided/Refused
Conservative
Moderate
Liberal
Don't know/Undecided/Refused
65+
50-64
30-49
18-29
Female
Male
Don't know/Undecided/Refused
$100K+
$50K-LT $100K
LT $50K
Don't know/Undecided/Refused
Coll Grad
Some Coll
H.S. Grad
No H.S.
Don't know/Undecided/Refused
Non-White
White
-10
-5
0
5
Change in % After Reweighting


NC: Effect of Reweighting by Vote and Mode by County Strata
Rep
Lean Rep
Ind (no lean)
Lean Dem
Dem
Don't know/Undecided/Refused
Conservative
Moderate
Liberal
Don't know/Undecided/Refused
65+
50-64
30-49
18-29
Female
Male
Don't know/Undecided/Refused
$100K+
$50K-LT $100K
LT $50K
Don't know/Undecided/Refused
Coll Grad
Some Coll
H.S. Grad
No H.S.
Don't know/Undecided/Refused
Non-White
White
-0.10
-0.05
0.00
0.05
Change in % After Reweighting

PA: Effect of Reweighting by Vote and Mode by County Strata




WI: Effect of Reweighting by Vote and Mode by County Strata
Rep
Lean Rep
Ind (no lean)
Lean Dem
Dem
Don't know/Undecided/Refused
Conservative
Moderate
Liberal
Don't know/Undecided/Refused
65+
50-64
30-49
18-29
Female
Male
Don't know/Undecided/Refused
$100K+
$50K-LT $100K
LT $50K
Don't know/Undecided/Refused
Coll Grad
Some Coll
H.S. Grad
No H.S.
Don't know/Undecided/Refused
Non-White
White
-10
-5
0
5
Change in % After Weighting

Appendix H: The Charge to the Task Force on 2020 Pre-Election Polling

Charge to Task Force on 2020 Pre-Election Polling
Evans Witt / Dan Merkle
June 27, 2019

AAPOR is committed to a fair and data-based evaluation of how the pre-election polls perform in the 2020 elections in the United States. This analysis will build on the report of the Ad Hoc Committee on 2016 Election Polling, *An Evaluation of 2016 Election Polls in the United States.*

Election years present particularly high profile moments for survey research. This is a time when polls dominate the media and the accuracy of polls can be confirmed or refuted by the actual election outcome. Sometimes, elections are followed by a chorus of concerns about "a crisis in polling" because when polls are "off" it makes for better headlines than when they are on-target.

Pre-election polling is critical to our industry. Such polling can support the democratic process and it offers a very public opportunity to showcase the benefits, and weaknesses, of survey research. Therefore, understanding and being able to articulate the overall performance of pre-election polling, the changing methodologies being used, and the potential for variation in the accuracy of polls is vital for our industry.

To support the collection and dissemination of such information, AAPOR will convene a panel of survey research and election polling experts to conduct a review of the 2020 pre-election polls. The goals of the task force will be to:

1. Evaluate the accuracy of 2020 pre-election polling for both the primaries and the general election on the presidential race and other races. To facilitate this analysis, the task force should attempt to gather the necessary information for each poll on an on-going basis during the election season.
2. Where necessary, examine why specific polls and/or methodologies failed to correctly estimate support for major party candidates, whether at the national level or state level
3. Review variation by different survey methodologies
4. Identify significant differences between pre-election polls in 2020 and polling in prior election years including 2018, 2016 and 2014.
5. Create an archive of the data collected for the analysis for 2020 and for prior years, to enable AAPOR to continue this work in future years.

An appropriately sized task force will be formed by Oct. 1, 2019, to begin work on the project.

The task force will gather information on all publicly released polls as follows:

Required: All polls on presidential state primaries and state caucuses
All polls on the general election for president at the national and state levels

Should be included: All general election polls on races for U.S. Senate
All general election polls on races for governor

Data on pre-election polls conducted with field dates from Jan. 1, 2020, to Nov. 3, 2020, will be eligible for inclusion in the analysis.

The task force should plan for a brief interim report on the performance of the polls in the presidential primaries by the end of July 2020. In addition, the task force will work with the Communications Committee and the Executive Committee on AAPOR statements about the polls during the 2020 campaign, as needed.

The draft report of the task force should be ready for review by Council at its January 2021 meeting. The final report should be ready for review by Council at its February 2021 meeting.

Once constituted, the task force will identify the steps needed to achieve the goals outlined here, including committee structure, data acquisition, planned analyses, report writing procedures, and any supplemental analyses, partnerships, or needed support.

###