DATA FOR *PROGRESS*

2020 Polling Retrospective

May 2021

Executive Summary

Since 2018 Data for Progress has helped prove the case for progressive policy with its best-in-class polling, while being on the forefront of innovative survey techniques. In the spirit of innovation, this report is the first of a series of analyses to rigorously assess polling's successes and failures over the past cycle while also offering concrete methodological improvements to address future errors.

In the week before the 2020 general election, Data for Progress conducted polling in 15 states with best-in-class accuracy. As we analyze our performance we've drawn three important conclusions:

- **Partisan Nonresponse and Activist Overrepresentation:** Conservative white voters are opting out of polling, while liberal voters are disproportionately opting in, creating an underlying bias in our respondent pools. We also have evidence that liberal partisan activists are systematically overrepresented in our surveys.
- **Geographic Heterogeneity in Respondents:** There are substantial differences between urban and rural white voters in terms of likelihood of voting Biden in 2020 or switching from Trump to Biden. Respondents living in zip codes which display the most firm Trump support are less likely to respond to polls — even when you control for their partisanship and other demographics.
- **Geographic Heterogeneity in Electoral Performance:** Biden overperformed House Democratic candidates in heavily white suburban and urban counties, and underperformed them in rural counties that were majority non-white.

DFP Polling during the Presidential Cycle

Data for Progress published[1] polls in 15 states, fielded Oct 27 - Nov 1 2020, using web panel and SMS respondents. Overall, our performance this presidential cycle compares favorably to other pollsters in both volume and accuracy, as shown in Figures 1 and 2. While using the same data as FiveThirtyEight's analysis,[2] Figure 2 is limited to only state polls in the two weeks preceding the election, with a floor of four polls rather than 10. This allows us to compare a broader set of pollsters as voters align shortly before the election. Our analysis excludes national polling which, while informative, has little bearing on the presidential results due to the Electoral College.

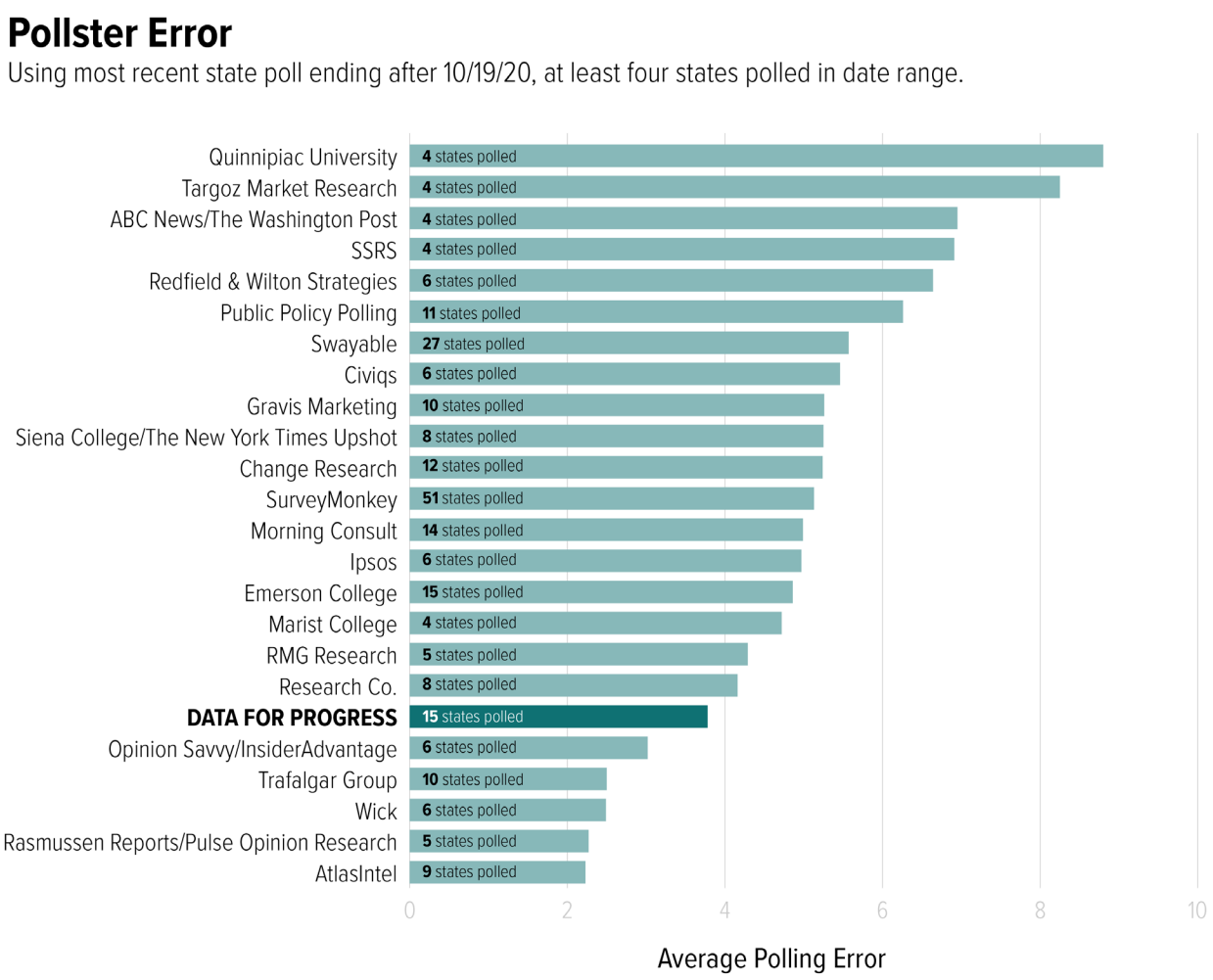
FIGURE 1
FINAL DFP PRE-ELECTION POLL MARGIN VS ACTUAL

	AL	AZ	CO	FL	GA	IA	KS	MN	MS	NC	NV	PA	SC	TX	VA
Actual	-25	0	13	-3	0	-8	-15	7	-16	-1	2	1	-12	-5	10
DFP	-20	3	12	3	2	-2	-14	8	-14	2	7	7	-9	1	11
Diff	5	3	-1	6	2	6	1	1	2	3	5	6	3	6	1

Traditional polling methods continue to struggle in terms of accuracy[3] and cost. To maintain accuracy, pollsters will have to rigorously adapt methodology on an ongoing basis, as they cope with declining response rates and difficulty getting good samples. This document describes lessons we've learned from this presidential polling and synopsizes areas for further research and iteration.

FIGURE 2

DFP'S PRESIDENTIAL POLL ACCURACY WAS BEST AMONG NON-GOP-ALIGNED POLLSTERS. DATA COURTESY OF FIVETHIRTYEIGHT



This chart measures the average polling error, the difference between a poll's topline results and the electoral outcome

DATA FOR ***PROGRESS***

Partisan Nonresponse

PARTISAN REPRESENTATION IN SURVEYS

The first issue that became clear throughout our polling this cycle is that there is substantial heterogeneity in how voters respond to polls, and the rates at which they respond. In addition to partisanship, we see significant differences in response rate and voting pattern by two major factors: education and geography. Our research into this effect originated with our SMS response rates, since lower SMS responses limit our polling respondent pools. From there, we took a larger look at factors that could explain response bias across both SMS and web polls, particularly focusing on geographic factors. To ensure our results were concrete, we validated these findings against 2020 election results.

SMS outreach to voters is one of the two primary ways Data For Progress contacts its respondents, and allows us to match them to the voter file, increasing our confidence that they are in fact voters.

We see two partisan effects among voters who respond to our polls:

- Democrats respond to polls at much higher rates than we expect based on the general population, so highly partisan liberals make up a larger proportion of respondents than they do in the general electorate.
- Republicans respond at somewhat lower rates than we expect, meaning that extreme and moderate conservatives make up a smaller proportion of respondents than they do in the general electorate.

As demonstrated by Figure 3 below, liberal voters, especially in urbanized areas, were overrepresented, while rural white voters were underrepresented in our polling. This effect is largely limited to white voters: both because they make up the vast majority of poll respondents, and because voters in minority communities tend to vote overwhelmingly for Democrats.

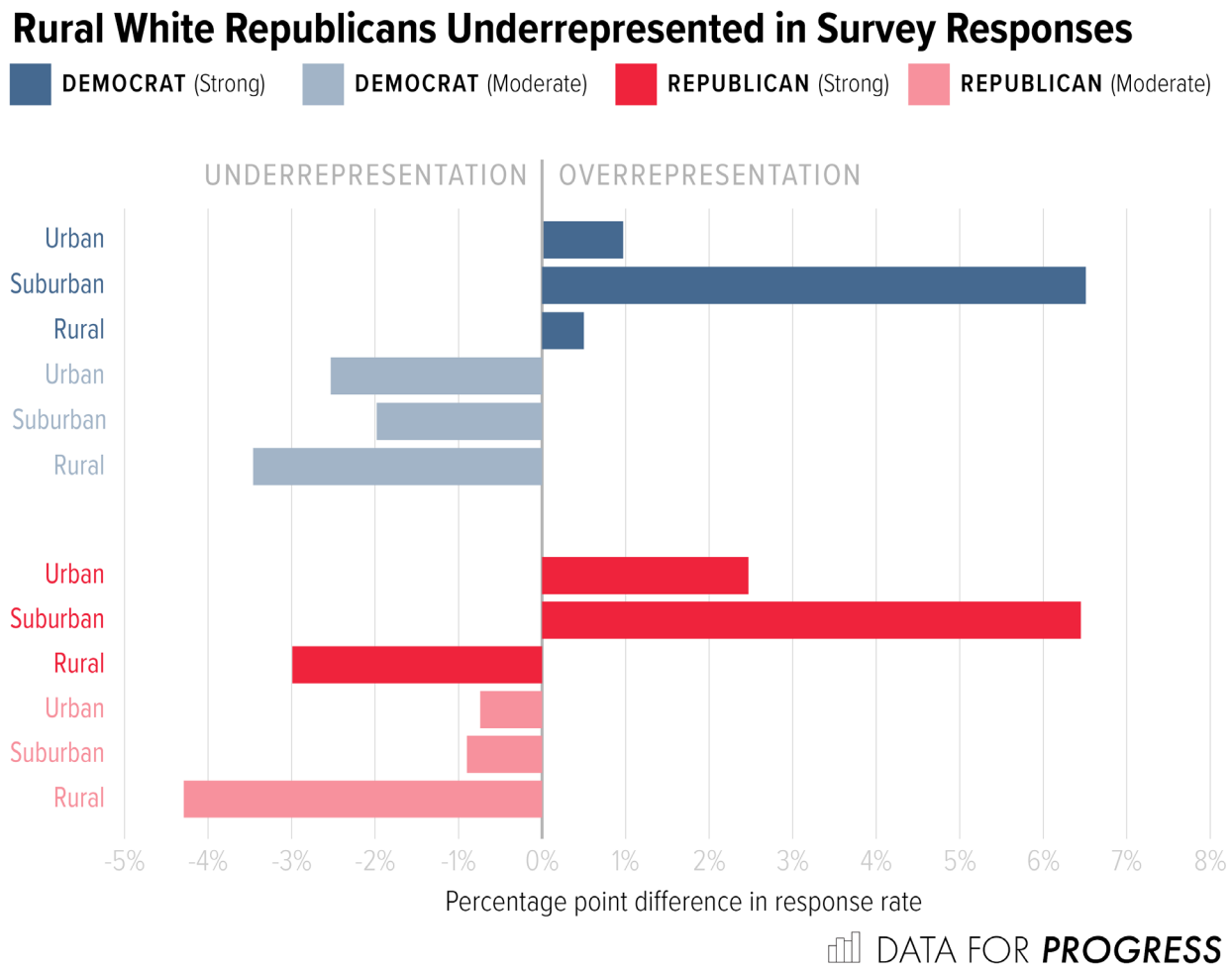
For this analysis we rely on a third-party partisanship score, a modeled propensity that a given voter will identify as a Democrat, as a proxy for partisanship on the x-axis. Using the score, we are able to compare the distribution of the partisanship score among respondents to the general population.

The two most striking examples of partisan under-response were Texas and Iowa. In Texas, middle and right partisans were underrepresented by 10 percentage points, while strong democratic partisans were overrepresented by 10pp[4]. In Iowa we saw the most dramatic difference: conservative voters underrepresented by 10pp, paired with a near 20pp overrepresentation of the most liberal respondents.

VOTER SMS REPLIES

Instead of opting out of polling, some voters reply to our SMS invitations — many of which feature expletives or scorn (those who reply with expletives are left out of our samples). After linking these respondents to the voter file, we can observe that conservative white voters account for the bulk of replies. To lend additional credibility to the hypothesis that some conservatives are actively opting out of polling, SMS responses frequently use expletives and mention a variant of either “Trump,” “Pence,” or “Maga.”

FIGURE 3



LIBERAL ACTIVIST OVERREPRESENTATION

In addition to our SMS analysis, we found heterogeneity among liberal survey respondents in both our SMS and web panel respondent pools. Using a third party activist score, which ranks voters by their propensity for activism (e.g. volunteering for a campaign) we can conduct similar representation analysis to assess if activists are overrepresented when compared with the voting population. We find, across geographic density, that liberal activists are overrepresented by as much as 10pp, and that voters with low activist scores are underrepresented by as much as 15pp.

This finding is complemented by ecological inference analysis[5] conducted by our team after the 2020 election in Georgia, using precinct-level election results disaggregated to census blocks to estimate the actual support for Biden among specific demographic groups. This methodology allows us to compare our polling with estimated actual support among demographic groups, demonstrating that we overestimated support for Biden among women and white college-educated voters, groups that also exhibit high rates of activism. While it is true that activists are an important part of the electorate, their overrepresentation will skew polling toplines.

While Republicans are opting out of polling, Democratic activists are disproportionately overrepresented in our surveys, contributing to the consistent overestimation of Biden support. While not conclusive, our research on response rates and responses allows us to account for response biases through weighting and improved sampling. A discussion of our improved weighting method is described in the next section.

FIGURE 4

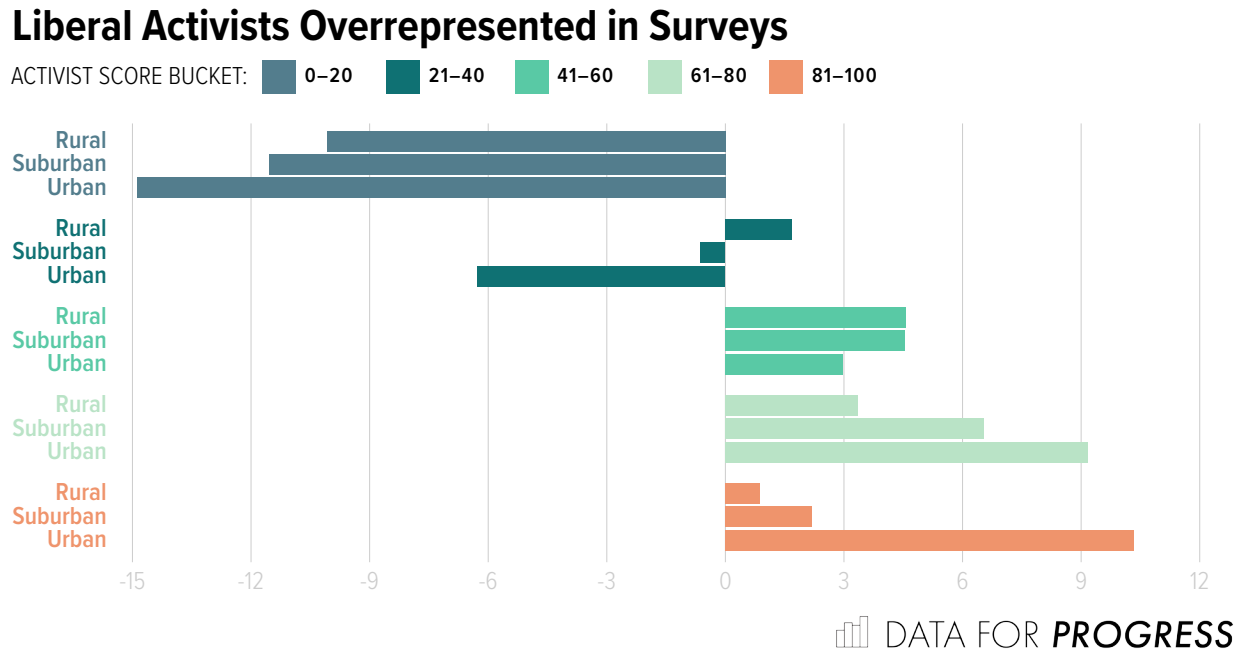
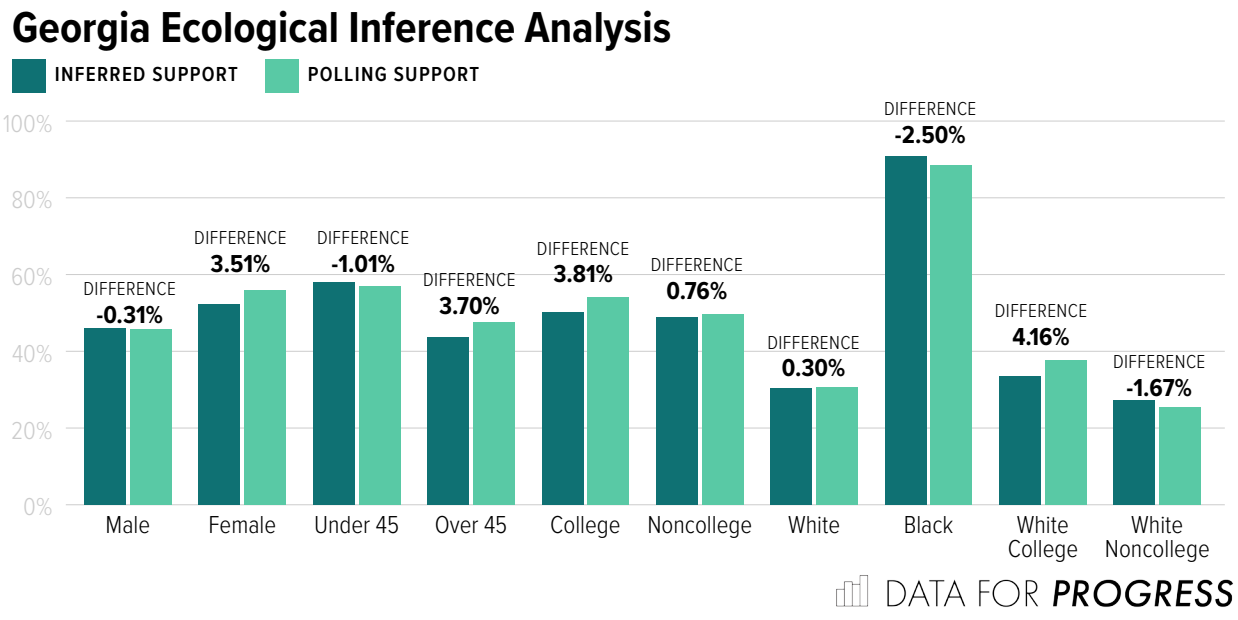


FIGURE 5



Geographic Heterogeneity in Respondents

We know that likelihood of responding to polls is related to individual partisanship and political beliefs[6], but in theory, if the most significant factor is partisanship, we can simply sample far more people we believe to be Republicans until we get enough, or upweight Republican respondents. This is a fundamental principle of polling: we replace uncontactable individuals with contactable individuals within demographic categories, and if the sample is small, give them greater weight in the survey. This method fails, however, when our ability to contact an individual is related to factors of interest.

In our research on both election results and a large database of survey responses, we find that there is significant heterogeneity in Biden support among white Trump 2016 voters based on political, economic, and demographic characteristics of the respondent's zip code.

This means that "white, Trump 2016" voters cannot be treated as a uniform block when sampling for polling. In fact, response rates for voters who live in zip codes that didn't swing between 2016 and 2020 are notably lower than response rates among similar voters in zip codes that swung more. We believe this effect is one of the reasons polling significantly overestimated the rate that voters would switch from Trump to Biden in 2020.

Our research has surfaced dozens of zip code characteristics which correlate with support for Trump in 2016 and 2020. For example, the percentage of college educated people in a person's zip code was a stronger predictor of Biden support than a person's individual education level.

While these findings might seem counter-intuitive, political science literature shows that the strongest Democratic and Democratically-trending areas are characterized by higher population density, higher concentrations of college-educated people, more diversity, and growing economies that are based on professional and waged service work. In contrast, the areas which are the most dedicated to Trump typically have a lower population density, higher concentrations of noncollege whites, and have economies that depend more on goods-producing sectors, often in declining post-industrial regions.[7]

To simplify this analysis, we reduced the various zip code characteristics into a single number that summarizes the odds that white voters would switch from Trump to Biden in 2020.[8] Below, we have binned this factor into population-weighted quantiles; higher quintile zip codes involve higher rates of Trump-to-Biden vote-switching, while lower quintile zip codes involve lower rates.

In Figure 6, we show unweighted Biden support along each quantile which demonstrates heterogeneity within white voters and across zip codes. Figure 7 illustrates the average weight of the respondents in each group and quantile, where weights above 1 indicate upweighting and those below 1 indicate downweighting.

FIGURE 6

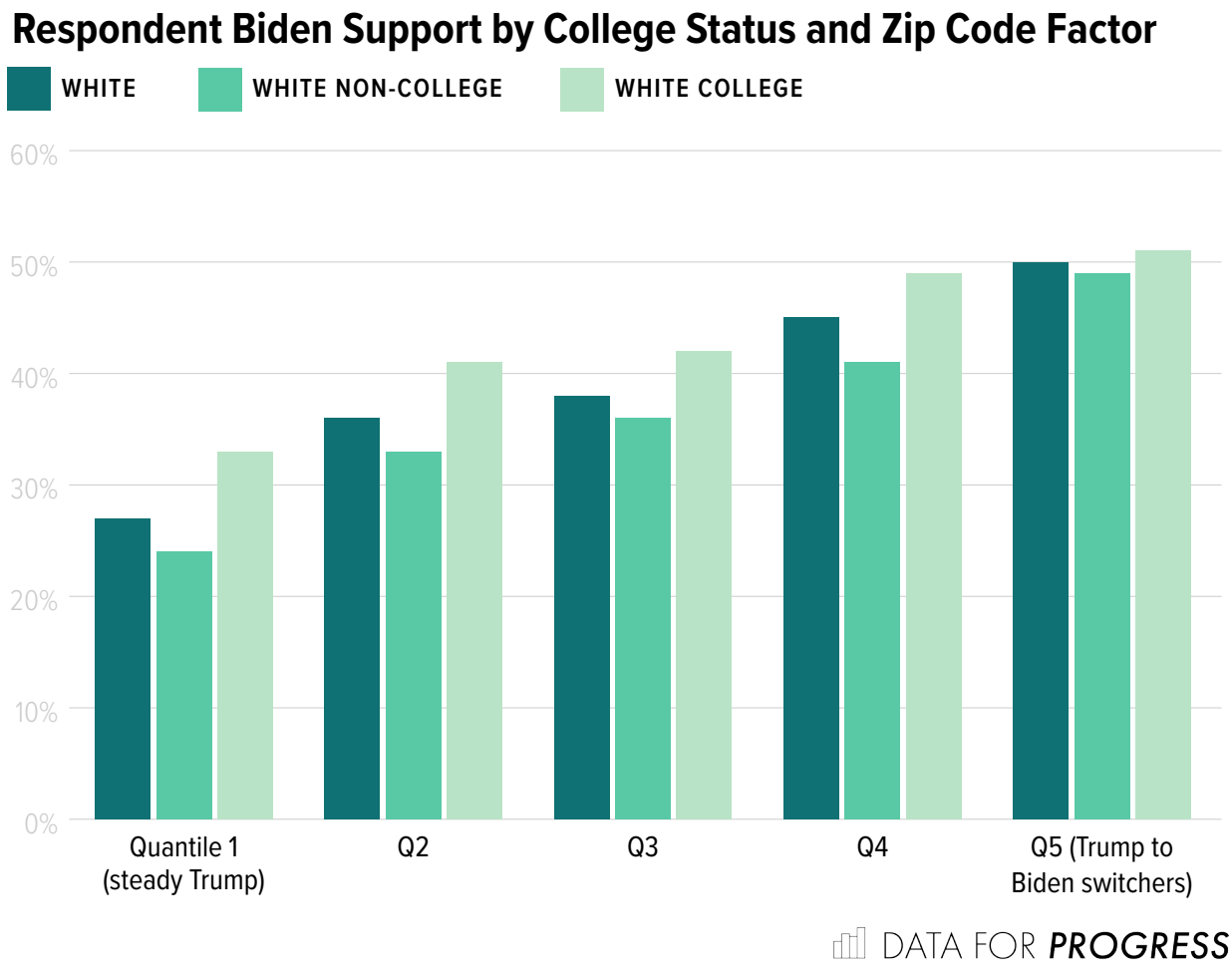
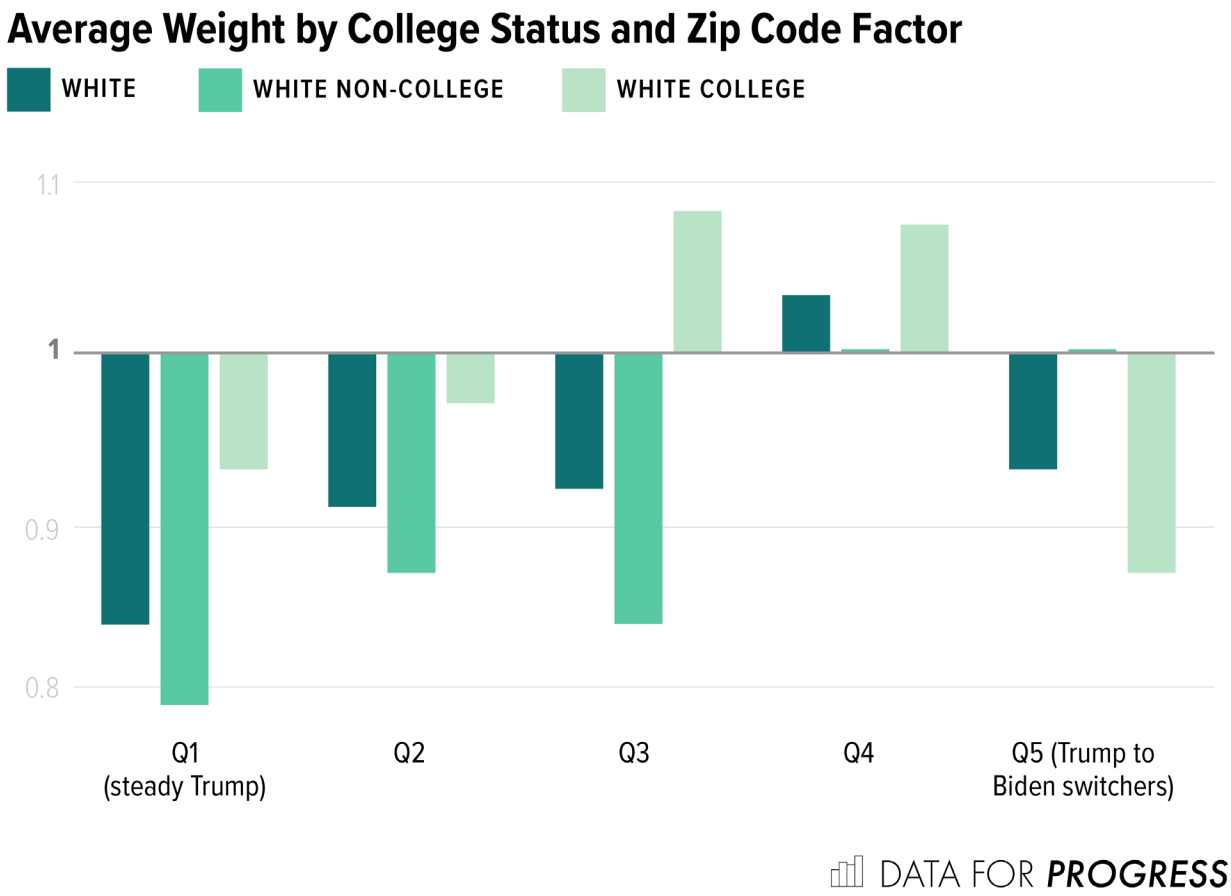


FIGURE 7



These results show two major things:

- There is significant heterogeneity within the white college/white non-college demographic groups, with Biden support within both groups being heavily influenced by the characteristics of their zip code.
- Our system was systematically downweighting voters from the most consistently pro-Trump zip codes, even within demographic groups.

This result highlights the inadequacy of the college-noncollege divide in fully capturing the dynamics of Trump support. In fact, we see that white college voters in Trump's strongest zip codes are extremely conservative, and white noncollege voters in Trump's weakest areas support Biden at rates higher than the national average for white college voters.

This is compelling evidence of a nonresponse mechanism that is biasing poll results, and we expect that geographic polarization will only intensify. This is not something that demographic weighting alone can fix. In fact, we were not only undersampling voters from the strongest Trump zip codes, but our weighting system was further down-weighting them.

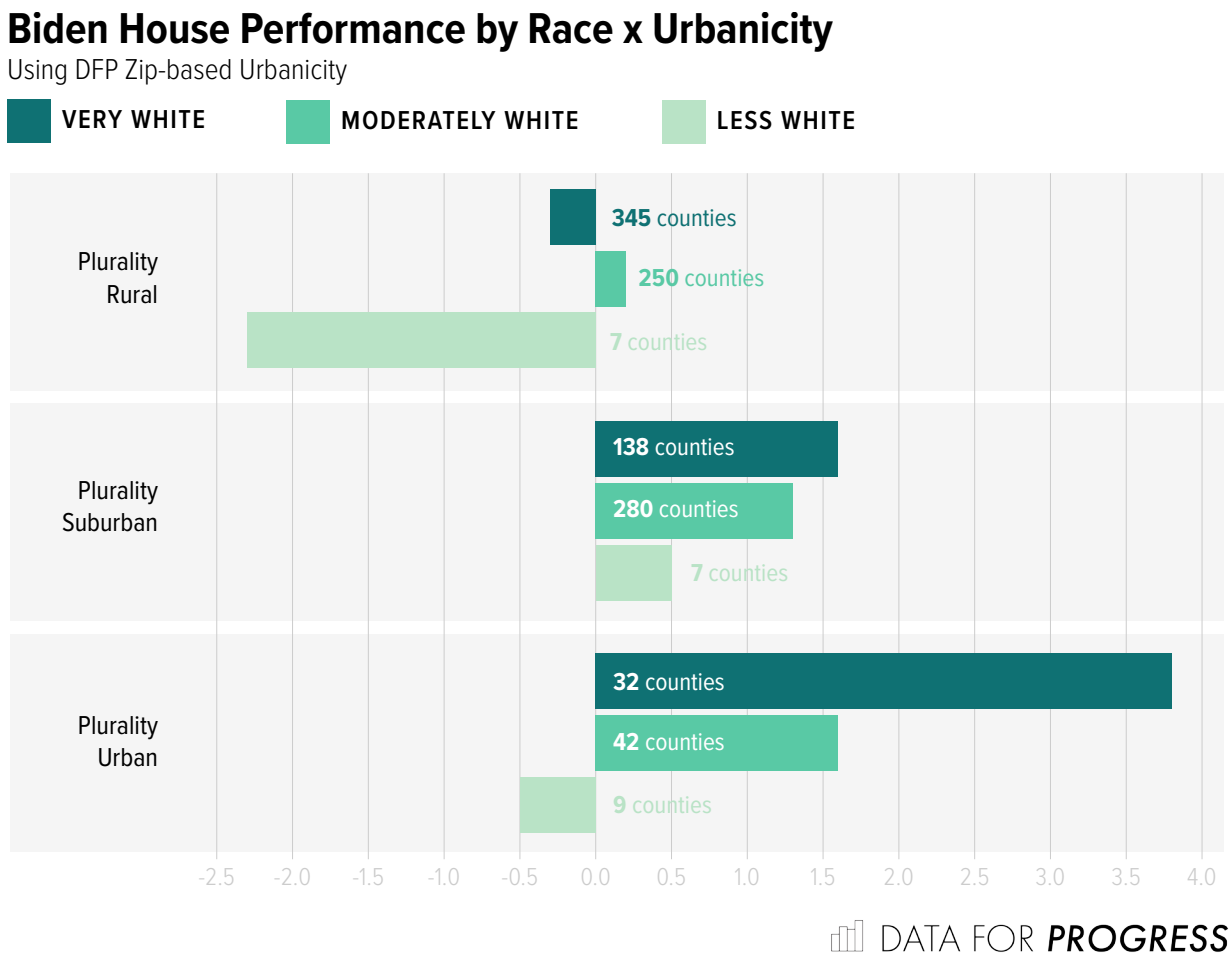
Geographic Heterogeneity in Electoral Performance

To validate the effects of geographic heterogeneity, we sought to examine the effects of Joe Biden on election results. One way to do so is to compare Joe Biden's performance against that of Democratic U.S. House candidates in the general election: any difference in the performance of those two is more likely the result of local factors than national partisan dynamics.

Specifically, we examined the relationship between Biden overperformance (relative to Democratic House candidates) to race and urbanicity, excluding uncontested House races. All analyses are conducted at the county-level and measure the difference in reported two-way vote share.

As Figure 8 shows, Biden overperformed House Democratic candidates in heavily white suburban and urban counties and underperformed in rural counties that were majority nonwhite. This further demonstrates that urbanicity is correlated with Biden vote share and exposed heterogeneity within white voters across geographies. We also found that rural areas had less defection from Trump between 2016 and 2020, with most of the Democratic presidential gains coming from urban or suburban areas. Combined with our findings about geographic heterogeneity in respondents, this presents both a theory of polling error and a general hypothesis of electoral trends from 2016 to 2020.

FIGURE 8



How We're Fixing It

Data for Progress polling performed well in the 2020 cycle, surveying more states more accurately than other progressive or media pollsters. That said, our initial analysis has identified three sources of error we're working to address:

- We've found that grouping respondents as "white college" or "white non-college" is no longer enough to capture the nuances of their voting behavior; there is significant variation within these groups. Our polling was systemically downweighting voters from the most consistently pro-Trump zip codes, missing geographic heterogeneity.
- White conservative voters are responding to polls at lower rates than we'd expect, creating a nonresponse bias. In fact, some conservatives are so passionate about not participating, they text us back.
- Liberal activists are overrepresented in our polling, and represent heterogeneities in the liberal respondent pool.

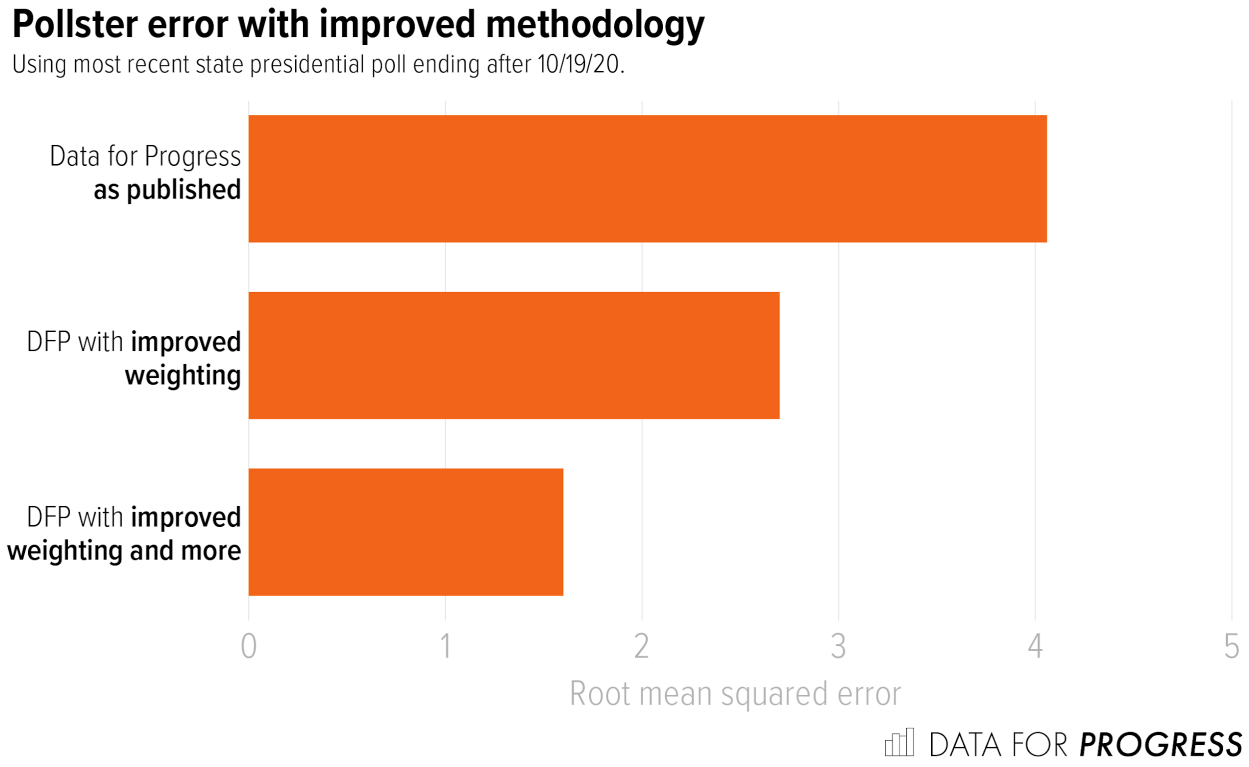
IMPROVED WEIGHTING

As we prepare for the 2021 election period, we have focused on improving our weighting using local characteristics — rooted in the ground truth of trusted public data sources like the U.S. Census. Some promising sources include census response rates, economic and employment indicators, and proxies for social trust.

We're applying findings about the importance of local geography to a zip code weighting scheme, which combines characteristics of respondents with the characteristics of their local geography. In fact, retroactively applying these weights to our 2020 pre-election surveys, we see that our new weighting techniques would have significantly decreased polling error (see Figure 9).

FIGURE 9

Our novel, improved weighting and methodology (estimated based on experimental results in sampling, quotas, recruitment messaging, in-survey messaging, and incentives) would have significantly lowered error



NONRESPONSE BIASES

As SMS-based and web panel survey research continues to evolve, we are experimenting with new techniques to increase sample representativeness:

- We are working to downweight and limit likely Democratic activists from our respondent pools using voter file scoring and survey questions to ensure we accurately mirror the voting population.
- We've begun A/B testing our phrasing in SMS introductions to encourage more conservative respondents to share their views, leveraging language used by popular conservatives to increase engagement.
- We are including content and adjusting our phrasing in surveys to maintain respondent participation and increase completion rates, while minimizing perceptions of bias in the survey itself.
- We are leveraging quotas and incentives to run surveys that oversample on partisanship, geography, and key demographic factors.
- We are continuing to monitor responses to SMS outreach to track partisan hostility (e.g. responses with "MAGA" or expletives), to track changes in nonresponse bias.
- Finally, we're beginning to explore new modes of engaging hard-to-reach populations: using cultural surveys and fielding them in new places, and including a small number of political questions at the end of the instrument.

This whitepaper is just the beginning. As we continue to pioneer new techniques, we're committed to sharing them and our findings with the progressive space. We believe that transparency and collaboration across the industry will help us all address these fundamental challenges with polling, ultimately empowering movements with better data.

In that spirit, we welcome feedback and collaboration! Please contact us if you have thoughts or would like to partner in the future.

AUTHORS

Colin McAuliffe | Colin is a co-founder of Data for Progress.

Johannes Fischer | Johannes is the Survey Methodology Lead at Data for Progress. Before DFP he was a data engineer with the DNC. He cares about polling, voter files, and dbt.

Charlotte Swasey | Charlotte is the Vice President of Data and Polling at Data For Progress. Before that, she worked on the Warren campaign on delegate math and resource allocation, as well as previously working in election forecasting and modeling.

Jason Katz-Brown | Jason is the CTO of Data for Progress. Before joining DFP, he was the Organizing Technology lead on Warren for President, where the team worked under a "SMALL TECH" mural and strove to empower every volunteer and voter in the fight for big, structural change.

Jason Ganz | Jason is the Chief of Staff with Data for Progress. His mission is making clean, accurate and accessible data on progressive issues available to everyone. Jason uses Python and SQL for analysis.

Gustavo Sanchez | Gustavo Sanchez is a Principal with Data for Progress where he helps manage the polling process from data collection through analysis. He serves as one of the group's methodological and technical process experts.

Sean McElwee | Sean McElwee is a Co-founder and Executive Director of Data for Progress.

ACKNOWLEDGEMENTS

Annie Wang contributed to analysis.

ENDNOTES

1. DFP Final Election Polling
2. https://fivethirtyeight.com/features/the-death-of-polling-is-greatly-exaggerated/
3. See "All that Glitters is not Gold".
4. Texas does not have mandatory party voter registration, so partisanship scores there tend to be less predictive than in states with party registration.
5. https://www.dataforprogress.org/blog/2019/5/9/inferring-vote-choice-without-a-survey
6. "The Mythical Swing Voter" and "How Pollsters Missed the 'Bowling Alone' Voters That Handed Trump the Presidency".
7. See "The Democratic Party's Suburban Shift Can Empower Progressives" and "The divide between us: Urban-rural political differences rooted in geography".
8. Formally, this zip code factor is the log odds of vote-switching conditional on zip code characteristics.

COVER PHOTO
Caleb Perez/Unsplash