US election polls: a quick postmortem

How did the 2020 US presidential election polls really do?
Ole J. Forsberg gives his assessment

The American Association for Public Opinion Research (AAPOR) is expected to produce a report early this year that explores the strengths and weaknesses of the polls in the 2020 US election cycle. The polls were criticised in some quarters immediately after the election, when it became clear that Donald Trump had done better than expected and that Joseph R. Biden Jr's margin of victory in the popular vote was not as large as anticipated.[1]

In preparation for this report, I wanted to provide some insight into the polls and some suggestions of my own for moving forward. Specifically, I hope to convince polling houses to use some type of model averaging – or even Bayesian methods – to reflect uncertainty in the voting population, and to encourage better explanations of poll results to the media and their readers.

Comparing 2016 and 2020

I expect that the AAPOR report, when published, will likely focus on the same three sources of error that were discussed in its May 2017 report covering the 2016 US election polls (bit.ly/3ihEYuH). According to that report, the 2016 polls underestimated Trump's eventual support because of (1) a failure to properly weight for the education level of respondents, (2) "shy Trump voters" outnumbering "shy Hillary Clinton voters" (either in response or non-response), and (3) a genuine shift in voter preference during the closing weeks of the campaign.

The first source of error, faulty weighting, is extremely important for polling houses to take seriously. While the number of US polling houses taking education level into consideration increased in 2020, the education characteristics of the voting population remain uncertain.

"Shy voters" – the second source of error – may be more myth than reality (53eig.ht/30NEb6R). But whether shy or not, there are some voters who either choose not to respond to polls, or who choose not to answer honestly when surveyed. Pollsters need to address this, either by asking additional questions to model respondent preference for those who choose not to say how they will vote, or by finding new ways to encourage the public to respond to legitimate polls, or even by using the non-response rate as an indicator of greater uncertainty in polling estimates.

I contend that the third source of error – a late shift in voter preference – is an error of interpretation, not of polling. The mistake happens in how we interpret a poll result such as "48% Biden, 44% Trump". Do we focus on the two-party vote and claim that Biden is ahead, or do we acknowledge that there is a sizeable portion of voters – 8% – who may only decide how to vote once in the polling booth? Clearly, the latter interpretation is more appropriate, but it makes for a less straightforward story, so these undecided voters tend to be overlooked in media reports.

Missing data

The majority of polls in the 2020 election cycle contained just three response options for those asked about their intended vote: "Biden", "Trump", and "undecided". The implied fourth option was "I refuse to take this poll" – and about 90% of people chose that "option" when contacted by a polling house (response rates were below 10%).

Taken together, these non-respondents and the undecideds mentioned earlier constitute a huge amount of missing data about voting intention. Ignoring these missing data leads to false precision in the polls' assessment of the state of the election.

While some undecided voters ultimately will not vote, many will eventually decide between the two candidates. This increases the uncertainty in polling estimates beyond what is reported in terms of confidence intervals and margins of error. As a result, when those late-deciding voters finally vote, polls may look very wrong.

To illustrate this point, compare the polls in the final two weeks of the 2020 election to the final election result (Table 1). In this sample of 174 polls, the actual Biden vote was within the polls' margins of error 85% of the time, while the actual Trump vote was within the polls' margins of error only 43% of the time. For the 57% of confidence intervals that missed Trump's actual vote, they were always too low, never too high – meaning that the polls consistently underestimated Trump's final vote. The 15% of confidence intervals that missed Biden's actual vote were roughly balanced between those that were too high and those that were too low. In other words, the polls tended to do a much better job of estimating the Biden vote than the Trump vote. But,

Table 1: Results from comparing candidate support levels in polls from the last two weeks of the US presidential election with the actual outcome of the election (vote share). Polls are a mix of state-level and national polls from a variety of polling houses, using a variety of methods.

Source	n	Confidence interval hits		Average miss (standard error)	
		Biden	Trump	Biden	Trump
All polls	174	85% (79% to 90%)	43% (35% to 50%)	–0.09	+2.41
Online only	23	96% (78% to 99%)	30% (13% to 53%)	–0.79	+2.21
Online + telephone	26	92% (75% to 99%)	54% (33% to 73%)	–0.78	+2.24
Telephone only	125	82% (74% to 88%)	42% (34% to 52%)	–0.18	+2.48
University	60	92% (82% to 97%)	27% (16% to 40%)	–0.10	+2.99
Non-university	114	82% (73% to 88%)	51% (41% to 60%)	–0.09	+2.10
Partisan	52	79% (65% to 89%)	75% (61% to 86%)	+0.62	+1.33
Non-partisan	122	88% (81% to 93%)	29% (21% to 38%)	–0.40	+2.87

Ole J. Forsberg is an assistant professor of mathematics-statistics and the director of the statistics program at Knox College in Illinois. He specialises in modelling and testing elections and is the author of *Understanding Elections through Statistics: Polling, Prediction, and Testing*.[2]
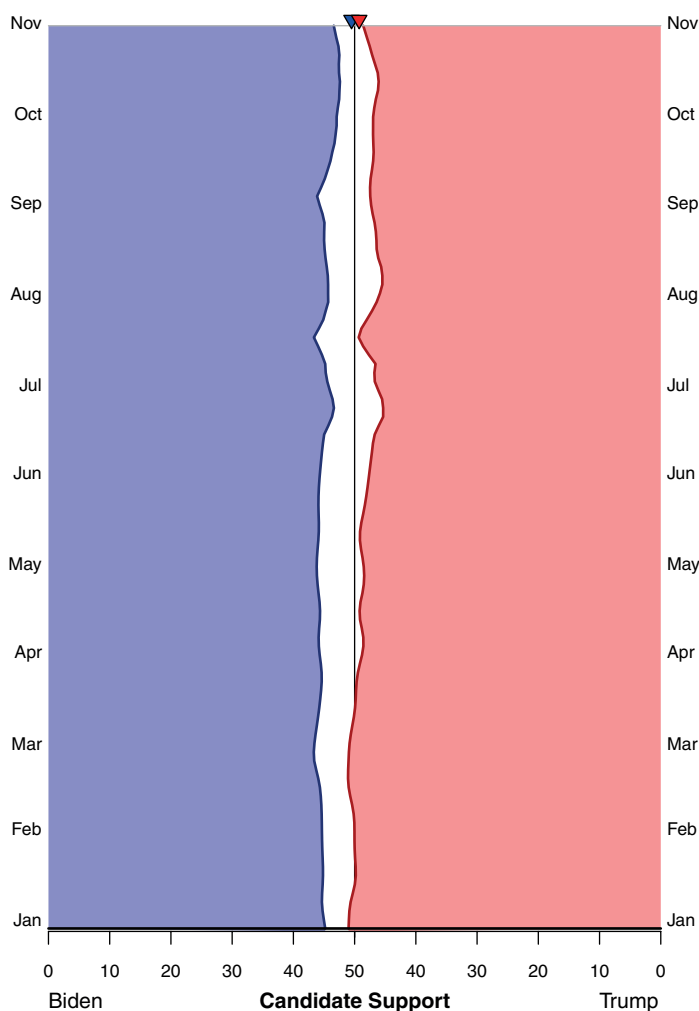
Figure 1: Estimated US presidential vote over time for Georgia, January to November 2020. The blue curve is the estimated support for Biden; the red curve, for Trump. The space between the two represents the proportion of undecided voters. The election results are indicated with the triangles at the top: Biden at 49.5%, Trump at 49.3%.

even so, they offered no clue as to how undecided voters would eventually vote.

Figure 1 illustrates this problem again, this time specifically for the state of Georgia. The left curve is the estimated support for Biden over time; the right, for Trump. The gap between the two curves represents the estimated proportion of undecided voters on any given day. That the election outcomes (triangles at the top) sit within this gap suggests that the polls did quite well in estimating each candidate's core support. But they failed in

estimating how the undecideds would break on election day.

The unknown population

Some will of course argue that it is not the job of polls to predict how people will vote, especially those who are undecided. Polls exist simply to offer a snapshot of how people say they intend to vote at a given point in time, based on a representative sample of the voting population. But here is where it gets tricky: the voting population does not exist until election day. There is a population of *eligible*

voters before election day, but not all eligible voters vote. This means that polling houses must estimate the characteristics of the voting population in order to recruit and weight their samples. They may use characteristics such as gender, political orientation, wealth, age and – yes – educational attainment. But because politics is not static, the characteristics used should be dependent on the election and its features. Crucially, pollsters will not know whether their samples are based on the right mix of population characteristics until after the election is won.

Currently, polling houses tend to settle on a single weighting scheme (weights assigned to each stratum in a stratified sample) and apply it to their raw data to achieve their final estimates. However, it would be more statistically sound for polling houses to acknowledge the uncertainty in the expected voting population and incorporate this into their estimates. This could be as simple as using several different "voting populations" to create several estimates of "voter support", for which pollsters then report the average. It could be as sophisticated as using Bayesian methods to place a prior distribution on the population strata and reporting the posterior mean and credible interval.

Personally, I favour the Bayesian solution because it provides a solid statistical structure for estimation and communication of results. Using Bayesian methods would force pollsters to acknowledge yet another source of uncertainty in their estimates and this may, in turn, encourage pollsters to be more modest with their results when communicating with the media. Such an approach may also help the media to better understand the inherent

uncertainty in poll results so that they can convey this to their readers, viewers, and listeners.

Communication

This leads me to what I think is a key lesson to be learned from the 2020 election polls. The end-user, the typical media consumer, tends not to have a solid understanding of statistics. Furthermore, they may not have the time to learn about statistics and what the poll numbers really mean. This places an additional burden on pollsters to ensure their results – their estimates, their uncertainties, *and* their meanings – are properly reported.

My view is that the polls, overall, did quite well. However, some media reports throughout the campaign failed to communicate what the polls were actually saying. Those same reports also failed to explain what polls are even *capable* of saying.

Polls provide a tantalising glimpse into the current state of some unknown future population. The presence of undecided voters adds to this uncertainty. If pollsters were to better convey this uncertainty and all that it means, it may lead the media to report polls differently, which may help to create reasonable expectations in future of what polls can and cannot tell us. ■

References
1. Tarran, B. (2020) US election called for Biden, as polls face criticism. *Significance*, **17**(6), 2.
2. Forsberg, O. J. (2021) *Understanding Elections through Statistics: Polling, Prediction, and Testing*. Boca Raton, FL: CRC Press.