

The New York Times

FiveThirtyEight

Nate Silver's Political Calculus

May 6, 2011, 9:00 pm

A Brief History of Primary Polling, Part III

By [NATE SILVER](#)

About a month ago, I published a pair of articles (see [Part I](#) and [Part II](#)) that focused on the predictive value of polls conducted in the early stage of the presidential primary process. Specifically, it looked at national polls conducted in January through June of the year before the primaries — for example, polls taken in January through June 1987 for the 1988 nomination races.

The articles were mainly descriptive, commenting on each primary cycle individually. Now it's time to look at the data in a more holistic way.

But first, some housekeeping. Since originally publishing the articles, Micah Cohen and I have done some additional quality control on our polling database. One significant ambiguity concerned cases where a pollster tested multiple scenarios within the same poll — for instance, versions of the 2008 Democratic primary field with and without Al Gore running. We had been handling these cases inconsistently. So instead, when these situations arise, we're now including each of the different scenarios in the averages, but assigning them a fractional weight. For instance, if a pollster tested four different versions of primary matchups with different sets of candidates running, we'd count them each at one-quarter weight.

In addition, we've combed through various online and offline resources and identified quite a few polls that we had been missing originally. And we've dropped the 1992 Republican primary from the analysis, because there were no polls conducted of that race until just a couple of months before the Iowa caucus.

The revisions to the database don't affect the big picture much (that is, the relative standing of the different candidates hasn't changed materially). Nevertheless, you can

find a table containing the updated data [here](#) — or download a spreadsheet version [here](#) (.xls file).

O.K., now for the fun part.

One way we can visualize this data is to compare a candidate's standing in the early polls to the share of the popular vote he eventually received throughout the primary process.

The correlation is far from perfect — there are candidates like Rudolph W. Giuliani who polled strongly but got no traction at all once voters began to cast their ballots, and there are candidates like Jimmy Carter who won their nominations despite having almost no early popular support. But the correlation is also far from zero (in fact, it's a moderately strong 0.72).

We can also place the candidates into different tiers based on how they placed in the early polling average. There is a fairly strong relationship between the candidates' polling and the number of states and votes they won during the primary process — as well as their chances of winning the nomination:

Of the seven candidates to poll at 35 percent or higher in the early polling average, five (or 71 percent) won their nominations, the exceptions being Hillary Rodham Clinton in 2008 and Ted Kennedy in 1980. Meanwhile, 31 percent of the candidates polling at 20 to 35 percent won their nominations, as did 13 percent of those polling at 10 to 20 percent, 4 percent of those polling at 5 to 10 percent, and so forth.

This relationship can be generalized by means of a [logistic regression](#) analysis:

A candidate polling at 1 percent at this stage of the primary race has about a 2 percent chance of winning his party's nomination, other factors being equal. A candidate polling at 10 percent, meanwhile, has only a 6 percent chance. But the probability begins to accelerate more rapidly once a candidate hits about 25 percent in the polling average — and once he reaches the low 30s, he becomes more likely than not to win the nomination.

One could take a variety of more sophisticated approaches with this data — for

instance, by accounting in some way for the relative standing of the candidates in addition to their raw numbers. Nevertheless, this underscores that it's simply quite wrong to suggest (as some [smart people have](#)) that early primary polls are meaningless. Instead, they have a reasonable amount of predictive power.

A more defensible hypothesis might be that one should account for any number of objective and subjective factors in addition to the national polls. It also might be the case that an expert could reliably identify candidates who were considerably stronger or weaker than suggested by their polling alone. I'm more dogmatic than most analysts about sticking to objective approaches, but I certainly would not argue that Donald J. Trump is more likely to win the Republican nomination than Tim Pawlenty, despite the fact that Mr. Trump now places well ahead of him in national surveys.

One factor that probably can improve one's predictions is accounting for name recognition. In addition to collecting the polling averages for past primary candidates, we also developed estimates of how widely a candidate's name was recognized in the early stages of a primary campaign. (Organizations like Gallup frequently [publish this data](#).)

Suppose, for example, that we divide the candidates into two cohorts: a high name recognition group (everyone with name recognition of 60 percent or higher at the beginning of the primary process) and a low name recognition one (everyone else).

The high-recognition candidates performed better on the whole; 10 percent of them won their nomination, as compared to 4 percent for the low-recognition group.

However, the low-recognition candidates performed better *relative to their polling average*.

A highly recognized candidate who finds himself in an inferior position in the polls is in a great deal of trouble. Of the 67 high-recognition candidates who were polling in the single digits, none won their nomination. In fact, only three of them won so much as a single state, and they averaged just 0.9 percent of the popular vote.

The odds are also against a low-recognition candidate who is polling in the single

digits. But those candidates at least have a fighting chance: 4 percent of them won their nominations and more than a dozen won at least one state, and they averaged 3.8 percent of the popular vote.

We can also run separate regression analyses on the high- and low-recognition groups.

Simply dividing a candidate's polling average by his name recognition is a reasonable approach. A candidate with 100 percent name recognition who is polling at 20 percent is roughly as likely to win his nomination as one with 50 percent name recognition who is polling at 10 percent.

In the fourth and final installment of the series, to be posted next week, we'll see what these approaches suggest to us about this year's Republican field.