Sep. 25, 2014

at

1:54 PM

How FiveThirtyEight Calculates Pollster Ratings

By Nate Silver

Filed under Polling

Get the data on GitHub

GitHub data at data/pollster-ratings

FacebookTwitterEmail

Illustration by Matt Chase

Illustration by Matt Chase

See FiveThirtyEight’s pollster ratings.

Pollster ratings were one of the founding features of FiveThirtyEight. I was rating pollsters before I was building election models. I was eagerly updating the ratings after every major batch of election results. I rated pollsters while walking two miles uphill ‚Ä¶ barefoot ‚Ä¶ in the snow. And then I got a little burned out on them. We last issued a major set of pollster ratings in June 2010 and made only a cursory update before the 2012 elections.

What happened? Well, when you publish a set of pollster ratings, people are understandably fixated upon how you’ve rated the individual polling firms: Is Pollster XYZ better than Pollster PDQ?

Naturally, we hope the pollster ratings can give you a better basis for understanding the polls as a news consumer. However, discussions about individual polling firms ‚Äî there are now more than 300 of them in our database ‚Äî can sometimes miss the point. I’m more interested in the big-picture questions. Are some pollsters consistently better than others, as measured by how accurately they predict election results? In other words, is pollster performance predictable? And if so, are a pollster’s past results the better predictor ‚Äî or are its methodological standards more telling?

The short answer is that pollster performance is predictable ‚Äî to some extent. Polling data is noisy and bad pollsters can get lucky. But pollster performance is predictable on the scale of something like the batting averages of Major League Baseball players.

Let me take that analogy a bit further. In baseball, there isn’t much difference in an absolute sense between a .300 hitter and a .260 hitter ‚Äî it amounts to getting about one extra hit during each week of the baseball season. Likewise, the differences in poll accuracy aren’t that large. We estimate that the very best pollsters might be about 1 percentage point more accurate than the average pollster over the long run. However, the average poll in our database missed the final election outcome by 5.3 percentage points. That means even the best poll would still be off by 4.3 points. It’s almost always better to take an average of polls rather than hoping for any one of them to ‚Äúhit a bullet with a bullet.‚Äù

What about the very worst pollsters? Well, we estimate that the absolute worst ones might introduce 2 to 3 points of error, as compared with average polls, based on poor methodology. That means that the worst polls are worse (further below average) than the best polls are good (above average). While there are intrinsic limits to how accurate any poll can be (because of sampling error and other factors), there is no shortage of ways to screw up.

But just as most baseball players hit somewhere around .260, most pollsters tend to be about average. Or at least, that’s the best guess we can make based on examining their past results. Poll accuracy statistics, like batting averages, take a long time to converge to the mean. You shouldn’t assume a polling firm is awesome just because it nailed the most recent election any more than you should mistake a shortstop who went 2-for-5 one day for a .400 hitter.

Nonetheless, when you aggregate results over a number of elections and the sample sizes become larger, you’ll find that there is some consistency in pollster performance.

Before we go any further, I’d encourage you to download the database of polls that we’ve used to construct the pollster ratings. We’re making it public for the first time. The database includes (with just a few minor exceptions that I’ll describe below) every poll conducted in the last three weeks of a presidential, U.S. Senate, U.S. House or gubernatorial campaign since 1998, along with polls in the final three weeks of presidential primaries and caucuses since 2000. Test everything out for yourself ‚Äî probably you’ll agree with some elements of our approach and disagree with others. Better yet, maybe you’ll discover a bunch of cool things that we hadn’t thought to look for. We think there should be more pollster ratings ‚Äî FiveThirtyEight shouldn’t have the last word on them.

Perhaps the simplest measure of poll accuracy is how far the poll’s margin was from the actual election result. For instance, if a poll had the Democrat ahead by 10 percentage points and she actually won by 3 points, that would represent a 7-point error.

In the table below, I’ve listed polling firms’ average error for elections from 1998 through 2007, and again for the same polling firms for elections from 2008 onward. (About half the polls in our database are from 2008 or later, so this is a logical dividing point.) I’ve restricted the list to the 28 firms with at least 10 polls in both halves of the sample.

As you can see, there’s a fair amount of consistency in these results; the correlation coefficient (where 1 is a perfect correlation and 0 is no correlation) is about 0.6. InsiderAdvantage and American Research Group were among the least accurate pollsters in both halves of the sample; polls from ABC News and The Washington Post (who usually conduct polls jointly) were among the most accurate in both cases. (ABC News, like ESPN and FiveThirtyEight, is owned by the Walt Disney Company.)

But there are a number of other things we’ll want to account for. In particular, we’ll want to know how much of the error had to do with circumstantial factors. For instance, polls of presidential primaries are associated with much more error than polls of general elections. This is a consequence of factors intrinsic to primaries (for instance, turnout is far lower) and mostly isn’t the pollsters’ fault. One more baseball analogy: Polling primaries is like hitting in Dodger Stadium against Clayton Kershaw, whereas polling general elections is like hitting in Coors Field.

How do we account for factors like these? It takes some work ‚Äî the balance of this article will be devoted to describing our process. This year, we’re publishing a variety of different versions of the pollster ratings that range from simple to more complex. If at any point you think we’ve made one assumption too many, you can take the exit ramp and use one of the simpler versions. Or you can download the raw data and construct your own.

Our overall method is largely the same as in 2010. That year, for the first time, we introduced a consideration of a poll’s methodological standards in addition to its past accuracy. We think the case for doing so has probably grown stronger since then, but you can find a number of versions of the pollster ratings based on past accuracy alone if you prefer them.

There are also a few things I’ve come to think about differently since 2010.

First, the case against Internet-based polls has grown much weaker in the last four years. At that time, the most prominent Internet pollster was Zogby Interactive (it has since been re-branded as JZ Analytics), which used a poor methodology and got equally poor results. But Internet penetration has increased considerably since then (it now exceeds landline telephone penetration) and a number of Internet-based polling firms with more thoughtful methodologies have come along. In 2012, the Internet-based polls did a little better than the telephone polls as a group (especially compared to telephone polls that did not call cellphones). There are still some reasons to be skeptical of Internet polls ‚Äî especially those that do not use probability sampling. But the FiveThirtyEight pollster ratings no longer include an explicit penalty for Internet polls1 as they did in 2010.

Second, it has become harder to distinguish a ‚Äúpartisan‚Äù poll. As I described earlier this month, FiveThirtyEight has been applying a more relaxed standard for what we define as ‚Äúpartisan‚Äù polls since 2012. The challenge in trying to use a more restrictive standard had been that there were too many borderline cases ‚Äî and we didn’t like having to make a lot of ad hoc decisions about which polls to include. Some polling firms, like Public Policy Polling, conduct polls on behalf of interest groups and campaigns but pay for others themselves. Blogs like Daily Kos and Red Racing Horses now sponsor polls. And in some cases, it isn’t clear who’s paying for a poll. Only the most unambiguously partisan polls ‚Äî those sponsored by candidates or by party groups like the Republican National Committee ‚Äî are excluded from the FiveThirtyEight forecast models.

But we still keep track of polls even when we don’t use them in our forecast models ‚Äî and their results are reflected in the pollster ratings. These polls are labeled with a partisan ‚Äúflag‚Äù in the database.2 The idea is that a polling firm ought to be held accountable for any poll it puts out for public consumption. If a polling firm releases biased and inaccurate polls on behalf of candidates, that will be reflected in its pollster rating ‚Äî even if it does better work when conducting polls on behalf of a media organization.

Our pollster ratings database also includes a couple of ways for you to track potential bias in the polls. The term bias itself refers to how much a polling firm’s results have erred toward one party or the other as compared against actual election results. House effect, by contrast, refers to how a firm’s results compare against other polls. If Pollster PDQ had the Democrat ahead by 5 points in an election where every other pollster had the race tied, it would have a Democratic house effect. But if the Democrat turned out to win by 10 points, PDQ would have a Republican bias as compared against the actual election results. As is the case for measures of poll accuracy, measures of bias and house effects can sometimes reflect statistical noise rather than anything systematic. But if they occur over dozens or hundreds of surveys, they should be a concern.

Third, we’re seeing clearer evidence of pollster ‚Äúherding.‚Äù Herding is the tendency of some polling firms to be influenced by others when issuing poll results. A pollster might want to avoid publishing a poll if it perceives that poll to be an outlier. Or it might have a poor methodology and make ad hoc adjustments so that its poll is more in line with a stronger one.

The problem with herding is that it reduces polls’ independence. One benefit of aggregating different polls is that you can account for any number of different methods and perspectives. But take the extreme case where there’s only one honest pollster in the field and a dozen herders who look at the honest polling firm’s results to calibrate their own. (For instance, if the honest poll has the Democrat up by 6 points, perhaps all the herders will list the Democrat as being ahead by somewhere between 4 and 8 points.) In this case, you really have just one poll that provides any information ‚Äî everything else is just a reflection of its results. And if the honest poll happens to go wrong, so will everyone else’s results.3

There’s reasonably persuasive evidence that herding has occurred in polls of Senate elections, presidential primaries and the most recent presidential general election. It seems to be more common among pollsters that take methodological shortcuts.

Paradoxically, while herding may make an individual polling firm’s results more accurate, it can make polling averages worse. There’s some tentative evidence that this is already happening. From our polling database, I compared two quantities: First, how accurate the average individual poll was; and second, how accurate the polling average was.4 I limited the analysis to general election races where at least three polls had been conducted.

From 1998 [TV] through 2007 [TV], the average poll in these races missed the final margin by 4.7 [QV] percentage points. The average error has been somewhat lower [TE] ‚Äî 4.3 [QV] percentage points ‚Äî in races from 2008 [TV] onward.

But the polling average hasn’t gotten any better ‚Äî if anything it’s gotten slightly worse [TE]. From 1998 [TV] through 2007 [TV], the polling average missed the final margin in an election by an average of 3.7 [QV] percentage points. Since [TE] 2008 [TV], the error has been 3.9 [QV] percentage points instead. LINE

So this is something we’re concerned about ‚Äî the benefit of aggregating polls together will decline if herding behavior continues to increase. This year’s pollster ratings introduce a couple of attempts to account for such behavior.

Now let’s get into the details ‚Äî what follows is a reasonably comprehensive description of how we calculate the pollster ratings.

Step 1: Collect and classify polls

Almost all of the work is in this step; we’ve spent hundreds of hours over the years collecting polls. The ones represented in the pollster ratings database meet three simple criteria:

They were conducted in 1998 or later;

They were conducted in the final three weeks of the campaign;

They were conducted in one of the following types of elections:

Presidential general elections;

Presidential primaries;

Senate elections;

Gubernatorial elections;

U.S. House elections.

Of course, it’s not quite that simple; a number of other considerations come up from time to time:

Sample sizes are sometimes missing from older polls. In these cases, we’ve estimated a poll’s sample size from its reported margin of error or from how many people a polling firm surveyed in other polls where the sample size was listed.5 As a last resort, we use 600 as a default sample size.

If a pollster listed results among likely voters and registered voters (or all adults), we list only the likely voter version in the database. Because the database covers the final three weeks of the campaign and almost all polling firms publish likely voter polls by that time, almost all polls in the database should be likely voter surveys.

When a pollster publishes multiple versions of the same survey (for example, versions of the poll with and without a third-party candidate included), FiveThirtyEight’s policy is to average the versions together. However, some of the polls in our database were taken from sources that may have followed different rules, so the treatment of these cases may be inconsistent.

Polls of special elections are included.

Polls of nonpartisan primaries (such as in Louisiana) are included.6

National polls for the presidential popular vote and the generic congressional ballot are included.7

The use of tracking polls is restricted to non-overlapping dates. For instance, if a firm’s final tracking poll was conducted on the Friday through the Sunday before an election, we wouldn’t also list the version that covered Thursday through Saturday.

Polls are included in the database even if they were not used in the FiveThirtyEight forecasts.8

A poll’s date as listed in the database reflects the median date the poll was in the field ‚Äî not the date the poll was released. For example, a poll conducted from Oct. 20 to Oct. 22 and released on Oct. 25 would have its date listed as Oct. 21.

Although in general all polls within the final three weeks of a campaign are included, there are minor exceptions in the case of the presidential primaries. No polls of the New Hampshire primary are included until after the Iowa caucus has been completed, and no polls of states beyond New Hampshire are included until New Hampshire has voted.9

Sources for the data include previous iterations of FiveThirtyEight, along with HuffPost Pollster, Real Clear Politics, PollingReport.com, the Internet Archive, and searches of Google News and other newspaper archives. They also include data sent to us by various polling firms ‚Äî however, we have sought to verify that such polls were in fact released to the public in advance of each election10 and that the pollster did not cherry-pick the results sent to us.

We’ve chosen 1998 as the cutoff point because there are multiple sources covering that election onward, meaning that the data ought to be reasonably comprehensive. Nevertheless, we’re certain that there are omissions from the database. We’re equally certain that there are any number of errors ‚Äî some that were included in the original sources, and some that we’ve introduced ourselves. We’re hoping that releasing the data publicly will allow people to check for potential errors and omissions.11

A big challenge comes in how to identify the pollster we associate with each survey. For instance, Marist College has recently begun to conduct polls for NBC News. Are these classified as Marist College polls, NBC News polls, NBC/Marist polls, or something else?

The answer is that they’re Marist College polls. Our policy is to classify the poll with the pollster itself rather than the media sponsor.

However, a few media companies like CBS News and The New York Times have in-house polling operations.12 Confusingly, media companies sometimes also act as the sponsors of polls conducted by other firms. Our goal is to associate the poll with the company that, in our estimation, contributed the most intellectual property to the survey’s methodology. For instance, the set of polls conducted earlier this year by YouGov for CBS News and The New York Times are classified as YouGov polls, not CBS News/New York Times polls.13

Polling firms sometimes operate under multiple brand names and add or subtract partners. Some cases are reasonably clear ‚Äî for instance, Rasmussen Reports is a subsidiary of Pulse Opinion Research, so polls marketed under each name are classified together. Other cases are more ambiguous; we’ve simply had to apply our best judgment about where one polling firm ends and another begins.

In previous versions of the pollster ratings, we included separate entries for telephone and Internet polls from the same company ‚Äî for instance, Ipsos conducts both types of polls and they’re listed separately in the database. This is becoming increasingly impractical as polling firms adopt mixed-mode samples (polls with Internet and telephone responses combined together) or otherwise fail to clearly differentiate one mode from the other. For now, we have grandfathered in preexisting cases like Ipsos and continued to list their Internet and telephone polls as separate entries. However, this will very likely change with the next major release of the pollster ratings database after the 2014 election.

Step 2: Calculate simple average error

This part’s really simple: We compare the margin in the poll against the election result and see how far apart they were. If the poll projected the Republican to win by 4 points and he won by 9 instead, that reflects a 5-point error. (Our preferred source for election results is Dave Leip’s Atlas of U.S. Presidential Elections.)

The error is calculated based on the margin separating the top two finishers in the election ‚Äî and not the top two candidates in the poll. For instance, if a certain poll had the 2008 Iowa Democratic caucus with Hillary Clinton at 32 percent, Barack Obama with 30 percent and John Edwards with 28 percent, we’d look at how much it projected Obama to win over Edwards since they were the top two finishers (Clinton narrowly finished third).

The database also includes a column indicating whether a poll ‚Äúcalled‚Äù the winner of the race correctly. But we think this is generally a poor measure of poll accuracy. In a race that the Democrat won by 1 percentage point, a poll that had the Republican winning by 1 point did a pretty good job, whereas one that had the Democrat winning by 13 was wildly off the mark.

Step 3: Calculate Simple Plus-Minus

As I mentioned, some elections are more conducive to accurate polling. In particular, presidential general elections are associated with accurate polling while presidential primaries are much more challenging to poll.14 Polls of general elections for Congress and for governor are somewhere in between.

This step seeks to account for that consideration along with a couple of other factors. We run a regression analysis that predicts polling error based on the type of election surveyed,15 a poll’s sample size,16 and the number of days17 separating the poll from the election.18

We then calculate a plus-minus score by comparing a poll’s average error against the error one would expect from these factors. For instance, Quinnipiac University polls have an average error of 4.6 [QV] percentage points. By comparison [CE], the average pollster, surveying the same types of races on the same dates and with the same sample sizes, would have an error of 5.3 [QV] points according to the regression. Quinnipiac [CV] therefore gets a Simple Plus-Minus score of -0.7 [QV].

This is a good score: As in golf, negative scores indicate better-than-average performance. Specifically, it means Quinnipiac polls have been 0.7 percentage points more accurate than other polls under similar circumstances.

A few words about the other factors Simple Plus-Minus considers: In the past, we’ve described the error in polls as resulting from three major components: sampling error, temporal error and pollster-induced error. They are related by a sum of squares formula:

\[Total\ Error=\sqrt{Sampling\ Error + Temporal\ Error + Pollster\text{-}Induced\ Error}\]

Sampling error reflects the fact that a poll surveys only some portion of the electorate rather than everybody. This matters less than you might expect; a poll of 1,000 voters will miss the final margin in the race by an average of only about 2.5 percentage points because of sampling error alone ‚Äî even in a state with 10 million voters.19 Unfortunately, sampling error isn’t the only problem pollsters have to worry about.

Another concern is that polls are (almost) never conducted on Election Day itself. I refer to this property as temporal (or time-dependent) error. There have been elections when important news events occurred in the 48 to 72 hours that separated the final polls from the election, such as the New Hampshire Democratic primary debate in 2008, or the revelation of George W. Bush’s 1976 DUI arrest before the 2000 presidential election.

If late-breaking news can sometimes affect the outcome of elections, why go back three weeks in evaluating pollster accuracy? Well, there are a number of considerations we need to balance against the possibility of last-minute shifts in the polls:

The overwhelming majority of elections do not feature important late-breaking developments. There will often be head-fakes and media-hyped ‚Äúgame changers,‚Äù but they rarely make much difference upon careful analysis.

Herding (see above) becomes more prominent in the final few days before an election. It’s fairly common for a pollster to publish some wild-seeming results ‚Äî which can affect media coverage of the campaign ‚Äî only to ‚Äúfall in line‚Äù with its final poll.

Some of the apparent movement in the polls in the late days of the election is probably artificial, reflecting response bias (voters for a certain candidate might be more likely to respond to polls after the candidate has a strong news cycle) and badly designed turnout models rather than genuine changes in public opinion.

‚ÄúElection Day‚Äù is something of a misnomer. Many states accept ballots by mail or provide for early voting; in the 2012 presidential election, about one-quarter of the votes nationwide were cast before Nov. 6.

Accounting for all polls in the final three weeks of the campaign increases the sample size.

Three weeks is an arbitrary cutoff point; I’d have no profound objection to expanding the interval to a month or narrowing it to two weeks, or to using a slightly different standard for primaries and general elections. But we feel strongly that evaluating a polling firm’s accuracy based only on its very last poll is a mistake.

Nonetheless, the pollster ratings account for the fact that polling on the eve of the election is slightly easier than doing so a couple of weeks out. So a firm shouldn’t be at any advantage or disadvantage because of when it surveys a race.

The final component is what we’ve referred to in the past as pollster-induced error; it’s the residual error component that can’t be explained by sampling error or temporal error. I’ve grown to dislike the term ‚Äúpollster-induced error‚Äù; it sounds more accusatory than it should. Certain things (like projecting turnout) are inherently pretty hard and it may not be the pollster’s fault when it fails to do them perfectly. Our research suggests that even if all polls were conducted on Election Day itself (no temporal error) and took an infinite sample size (no sampling error) the average one would still miss the final margin in the race by about 2 percentage points.

However, some polling firms are associated with more of this type of error. That’s what our plus-minus scores seek to evaluate.

Step 4: Calculate Advanced Plus-Minus

Earlier this year, House majority leader Eric Cantor lost his Republican primary to David Brat, a college professor, in Virginia’s 7th congressional district. It was a stunning upset, at least according to the polls. For instance, a Vox Populi poll had put Cantor ahead by 12 points. Instead, Brat won by 12 points. The Vox Populi poll missed by 24 points.

According to Simple Plus-Minus, that poll would score very poorly. We don’t have a comprehensive database of House primary polls and don’t include them in the pollster ratings, but I’d guess that such polls are off by something like 10 percentage points on average. Vox Populi’s poll missed by 24, so it would get a Simple Plus-Minus score of +14.

That seems pretty terrible ‚Äî until you compare it to the only other poll of the race, an internal poll released by McLaughlin & Associates on behalf of Cantor’s campaign. That poll had Cantor up by 34 points ‚Äî a 46-point error! If we calculated something called Relative Plus-Minus (how the poll compares against others of the same race) the Vox Populi poll would get a score of -22, since it was 22 points more accurate than the McLaughlin survey.

Advanced Plus-Minus, the next step in the calculation, seeks to balance these considerations. It weights Relative Plus-Minus based on the number of distinct polling firms20 that surveyed the same race, then treats Simple Plus-Minus as equivalent to three polls. For example, if six other polling firms surveyed a certain race, Relative Plus-Minus would get two-thirds of the weight and Simple Plus-Minus would get one-third.

The short version: When there are a lot of polls in the field, Advanced Plus-Minus is mostly based on how well a poll did in comparison to others of the same election. But when there is scant polling, it’s mostly based on a comparison to polls of the same type of election (for example, other presidential primaries).

Meticulous readers might wonder about another problem. If we’re comparing a poll against its competitors, shouldn’t we account for the strength of the competition? If a pollster misses every election by 40 points, it’s easy to look good by comparison if you happen to poll the same races. The problem is similar to the one you’ll encounter if you try to design college football or basketball rankings: Ideally, you’ll want to account for strength of schedule in addition to wins and losses and margin of victory. Advanced Plus-Minus addresses this by means of iteration (see a good explanation here), a technique commonly applied in sports power ratings.

Advanced Plus-Minus also addresses another problem. As I’ve mentioned, polls tend to be more accurate when there are more of them in the field. This may reflect herding, selection bias (pollsters may be more inclined to survey easier races; consider how many of them are avoiding the challenging Senate races in Kansas and Alaska this year), or some combination thereof. So Advanced-Plus Minus also adjusts scores based on how many other polling firms surveyed the same election. This has the effect of rewarding polling firms that survey races few other pollsters do and penalizing those that swoop in only after there are already a dozen polls in the field.

Two final wrinkles. Advanced Plus-Minus puts slightly more weight on more recent polls.21 It also contains a subtle adjustment to account for the higher volatility of certain election types, especially presidential primaries.22

Before we proceed to the final step, let’s pause to re-examine the results for the 28 polling firms we listed before, but this time using Advanced Plus-Minus rather than Simple Average Error.

There’s still a correlation ‚Äî although it’s somewhat weaker than before (the correlation coefficient is roughly 0.45 instead of 0.60). Accounting for the fact that American Research Group polls a lot of primaries makes the firm look somewhat less bad, for instance.

But pollster performance still looks to be predictable to some extent. As I’ll describe next, it’s more predictable if you look at a poll’s methodological standards in addition to its past performance.

Step 5: Calculate Predictive Plus-Minus

When we last updated the pollster ratings in 2010, I failed to be explicit enough about our goal: to predict which polling firms would be most accurate going forward. This is useful to know if you’re using polls to forecast election results, for example.

But that may not be your purpose. If you’re interested in a purely retrospective analysis of poll accuracy, there are a number of measures of it in our pollster ratings spreadsheet. For instance, you’ll find each pollster’s Simple Plus-Minus and Advanced Plus-Minus scores. The version I’d personally recommend is called ‚ÄúMean-Reverted Advanced Plus-Minus,‚Äù which is retrospective but discounts the results for pollsters with a small number of polls in the database.23

The difference with Predictive Plus-Minus is that it also accounts for a polling firm’s methodological standards ‚Äî albeit in a slightly roundabout way. In 2010, we looked at whether a polling firm was a member of the National Council on Public Polls (NCPP) or a supporter of the American Association for Public Opinion Research (AAPOR) Transparency Initiative.24

One other thing I was probably not clear enough about in 2010 was that participation in these organizations was intended as a proxy variable for methodological quality. That is, it’s a correlate of methodological quality rather than a direct measure of it.25 Nevertheless, polling firms that participated in one of these initiatives tended to have more accurate polls prior to 2010. Have they also been more accurate since?

Yes they have ‚Äî and by a wide margin. The chart below tracks the performance of polling firms based on whether they were members of NCPP or the AAPOR Transparency Initiative as of June 6, 2010, when FiveThirtyEight last released a full set of pollster ratings.26

From 1998 through 2009, the average poll from an AAPOR/NCPP polling firm had an error [CV] of 4.6 percentage points [QV], compared with [CE] an average error [CV] of 5.5 percentage points [QV] for firms that did not participate in one of these groups. While this difference is highly statistically significant, it isn’t that impressive. BAR

The reason is that we evaluated participation in AAPOR/NCPP only after the fact. Perhaps polling firms with terrible track records didn’t survive long enough to participate in AAPOR/NCPP as of June 2010, or perhaps AAPOR/NCPP didn’t admit them.

What is impressive is that the difference has continued to be just as substantial since June 2010.

In the general election in November 2010, polls from firms that had participated in AAPOR/NCPP as of that June [CV] were associated with an average error of 4.7 percentage points [QV], compared with [CE] 5.7 percentage points [QV] for those that hadn’t [CV]. BAR

And throughout 2012 (including both the presidential primaries and the general election), the AAPOR/NCPP polls [CV] were associated with an average error of 4.0 percentage points [QV], versus 5.2 points [QV] for nonparticipants [CV]. BAR

For clarity: The 2010 and 2012 results are true out-of-sample tests. In the chart above, the polling firms are classified based on the way FiveThirtyEight had them in June 2010 ‚Äî before these elections occurred. In my view, this is reasonably persuasive evidence that methodology matters, at least to the extent we can infer something about it from AAPOR/NCPP participation.

This year, we’ve introduced a two-pronged test for methodological quality. The first test is similar to before: Is a polling firm a member of NCPP, a participant in the AAPOR Transparency Initiative, or does it release its raw data to the Roper Center Archive?27 And second, does the firm regularly call cellphones in addition to landlines? Each firm gets a methodological score between 0 and 2 based on the answers to these questions.

Tracking which firms call cellphones is tricky. We’ve done a reasonably extensive search through recent polls to see whether they document calling cellphones. However, we do not list a polling firm as calling cellphones until we have some evidence that it does. There are undoubtedly some false negatives on our list; we encourage polling firms to contact us with documentation that they’ve been calling cellphones.28

So let’s say you have one polling firm that passes our methodological tests but hasn’t been so accurate, and another that doesn’t meet the methodological standards but has a reasonably good track record. Which one should you expect to be more accurate going forward?

That’s the question Predictive Plus-Minus ratings are intended to address. But the answer isn’t straightforward; it depends on how large a sample of polls you have from each firm. Our finding is that past performance reflects more noise than signal until you have about 30 polls to evaluate, so you should probably go with the firm with the higher methodological standards up to that point. If you have 100 polls from each pollster, however, you should tend to value past performance over methodology.29

One further complication is herding. The methodologically inferior pollster may be posting superficially good results by manipulating its polls to match those of the stronger polling firms. If left to its own devices ‚Äî without stronger polls to guide it ‚Äî it might not do so well.

My colleague Harry Enten looked at Senate polls since 2006 and found that methodologically poor pollsters improve their accuracy by roughly 2 percentage points when there are also strong polls in the field. My own research on the broader polling database did not find quite so large an effect; instead it was closer to 0.6 percentage points. Still, the effect was highly statistically significant. As a result, Predictive Plus-Minus includes a ‚Äúherding penalty‚Äù for pollsters with low methodology ratings.30

The formula for how to calculate Predictive Plus-Minus is included in the footnotes.31 Basically, it’s a version of Advanced Plus-Minus where scores are reverted toward a mean, where the mean depends on whether the poll passed one or both methodological standards.32 The fewer polls a firm has, the more its score is reverted toward this mean. So Predictive Plus-Minus is mostly about a poll’s methodological standards for firms with only a few surveys in the database, and mostly about its past results for those with many.33

As a final step, we’ve translated each firm’s Predictive Plus-Minus rating into a letter grade, from A+ to F. One purpose of this is to make clear that the vast majority of polling firms cluster somewhere in the middle of the spectrum; about 84 percent of polling firms receive grades in the B or C range.

There are a whole bunch of other goodies in the pollster ratings spreadsheet, including various measures of bias and house effects. We think the pollster ratings are a valuable tool, so we wanted to make sure you had a few more options for how to use them.

CORRECTION (May 21, 2016, 4 p.m.): An earlier version of this article included an incorrect date in the formula in Footnote 21. The date should be 1988, not 1998.