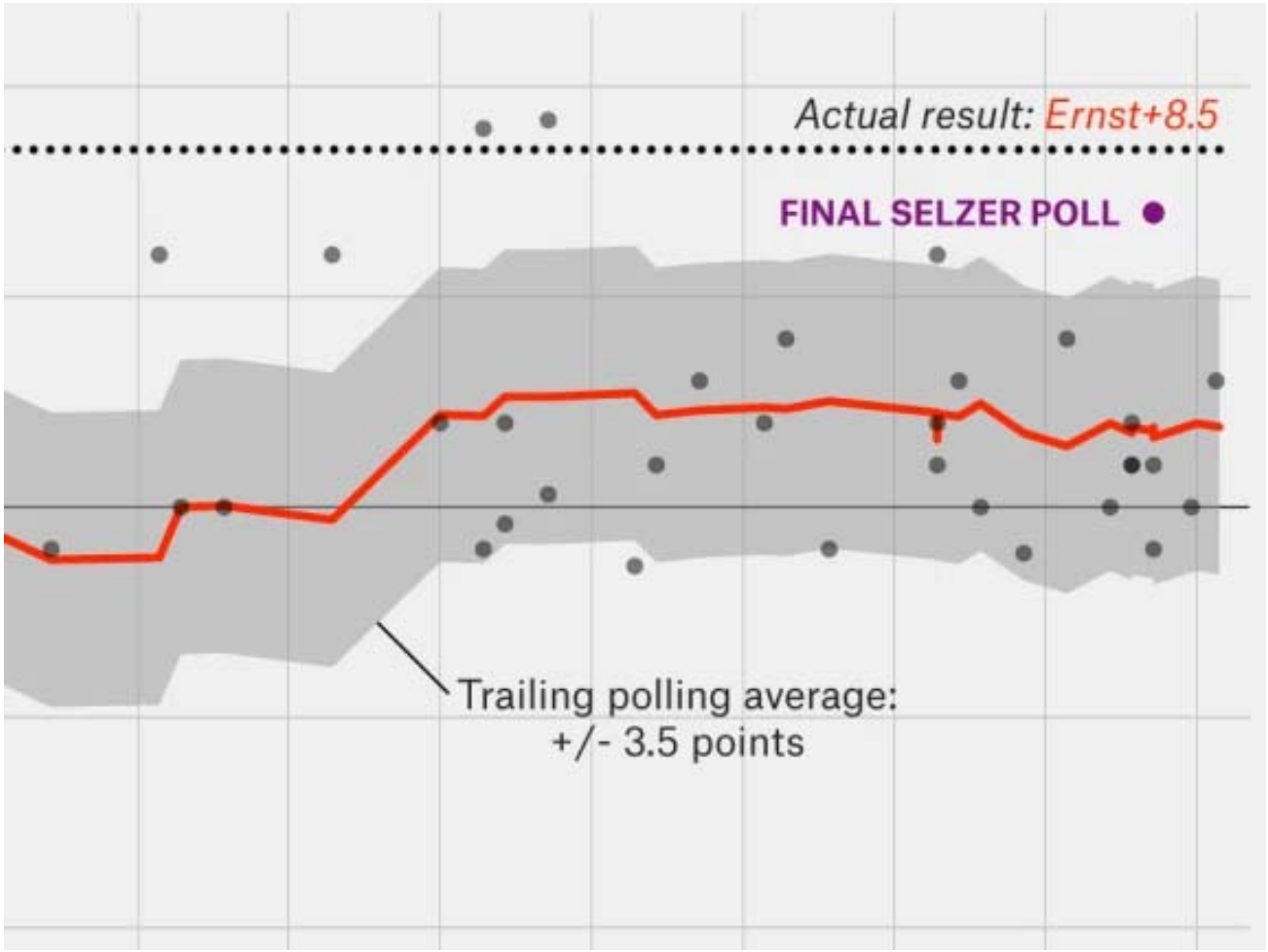
NOV. 14, 2014, AT 1:58 PM

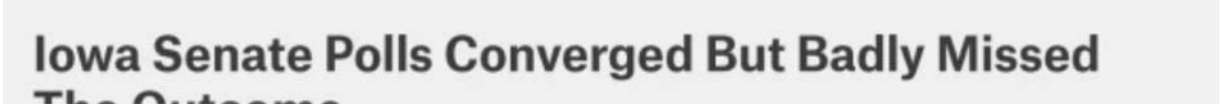
Here's Proof Some Pollsters Are Putting A Thumb On The Scale

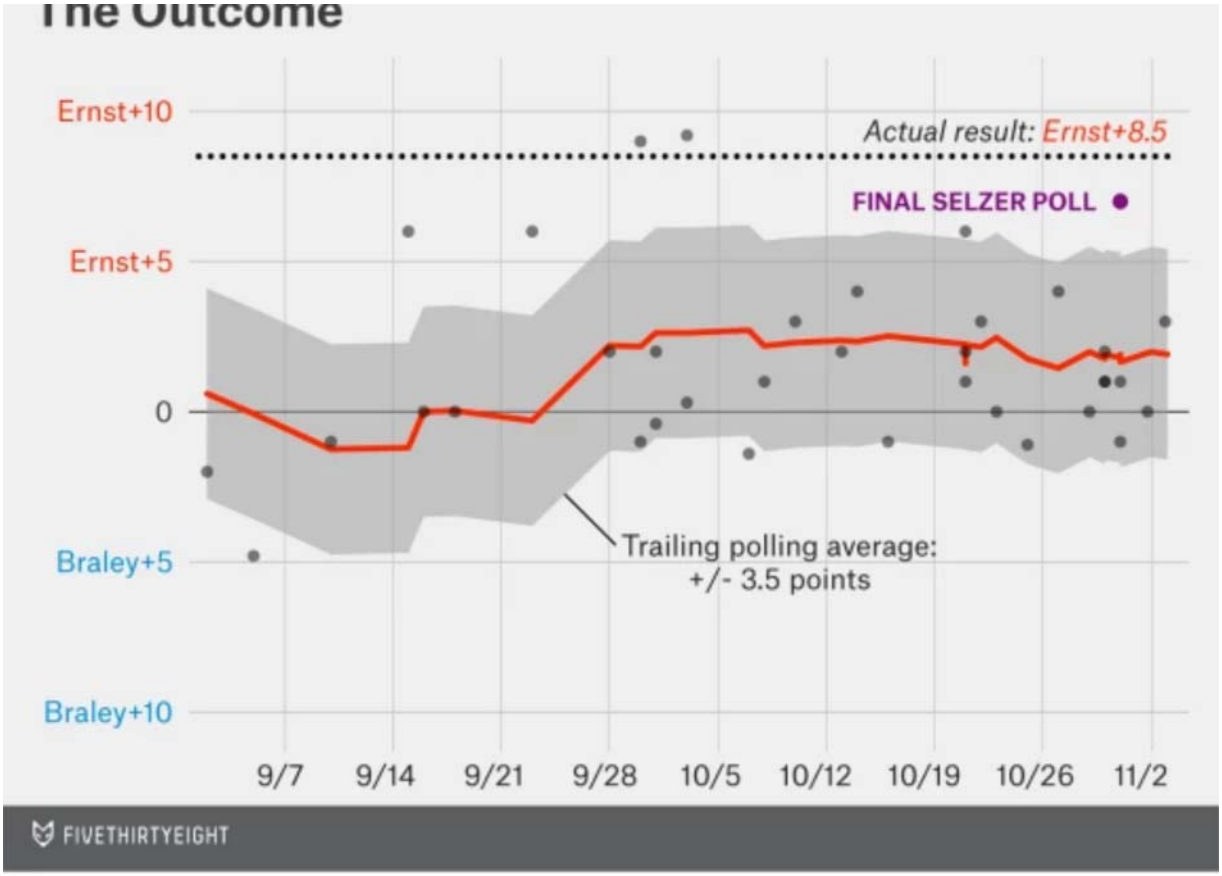
By Nate Silver

Filed under Polling



It's time to stop worrying about outliers and start worrying about inliers. Earlier this year, my colleague Harry Enten documented evidence of pollster "herding" — the tendency of polling firms to produce results that closely match one another, especially toward the end of a campaign.[1] What's wrong with the polls agreeing with one another? The problem is that it's sometimes a case of the blind leading the blind. Take a look at the polls conducted in this year's Senate race in Iowa, for example:





This chart depicts every likely voter poll conducted over the final nine weeks of the campaign and how each compared to the polling average at the time. (We'll get into more detail about how this polling average is calculated later on.) In September, when many voters (and pollsters) were just tuning into the race, there was plenty of diversity in the Iowa polls. A Loras College poll completed on Sept. 5 put Democrat Bruce Braley ahead by almost 5 percentage points against Republican Joni Ernst. Just 10 days later, Quinnipiac University completed a poll showing Ernst up by 6 points instead. That would soon be followed by a Selzer & Company poll for the Des Moines Register that had Ernst ahead by the same 6-point margin.

By the end of the campaign, however, the polls were in much stronger agreement. Twelve of the final 13 surveys had the race at somewhere between a 1-point lead for Braley and a 4-point lead for Ernst — a tight consensus suggesting a narrow edge for the Republican. Even Quinnipiac had Iowa tied in its final poll. The lone exception was Selzer's final poll for the Des Moines Register, which had Ernst up by 7 points — a result that pollster J. Ann Selzer would take an awful lot of grief about despite her stellar track record.

Better ignore that "outlier" poll from Selzer, right? Nope, not in this case. Ernst ended up winning by 8.5 percentage points. Most polls correctly identified Ernst as the

winner, but Selzer's poll was the only one in the final days to come close to her margin of victory.

It's not the inaccuracy of the polling average that should bother you — Iowa was one of many states where the polls overestimated how well Democrats would do — so much as the consensus around such a wrong result. This consensus very likely reflects herding. In this case, pollsters herded toward the wrong number.

In the Iowa chart, the shaded gray area represents the polling average plus or minus 3.5 percentage points. You can see how a number of polls fell outside the shaded area in September but that only the Selzer poll did in the final week or two of the race.

The 3.5 point range is important because it reflects sampling error: the intrinsic, unavoidable uncertainty introduced by taking a random sample of voters rather than surveying the whole population. Specifically, it's the standard error associated with a poll sampling 800 voters — the typical size of a Senate poll this year — in estimating the margin between the candidates. The standard error is not quite the same thing as the more familiar margin of error, but it gets at the same idea.[2] Whereas the "true" result[3] should fall within the margin of error 95 percent of the time, it should fall within the range established by the standard error about 68 percent of the time.

This necessarily implies that about 32 percent of results should fall *outside* the standard error.[4] As I said, sampling error is unavoidable — an intrinsic part of polling. If you've collected enough polls and don't find that at least 32 percent of them deviate from the polling average by 3.5 percentage points,[5] it means something funny — like herding — is going on.

In fact, 32 percent is an optimistic estimate. It accounts for sampling error alone and not the other sources of uncertainty in polling. It's hard to reach certain voters and hard to know who will turn out to vote, especially in midterm elections. News events can change the campaign after you've conducted your poll.

But sampling error alone produces considerably noisier polling than you might expect. The next table (click to expand) consists of a series of simulations where I've conducted "polls" by drawing random numbers from a normal probability distribution with a standard error of 3.5 percentage points. To make things more familiar, I've calibrated the numbers so they're distributed around an average that matches how the Senate polling looked late this year. In Iowa, for instance, the average has Ernst ahead by 2 percentage points and the standard error of 3.5 points is distributed around that

number.

Real Polling Data Is Noisy

Simulated polling results around a known mean, accounting for sampling error in an 800-voter survey

STATE	KNOWN MEAN	SIMULATION 1	SIMULATION 2	SIMULATION 3	SIMULATION 4	SIMULATION 5
Alaska	Sullivan +2	Tie	Sullivan +2	Sullivan +1	Begich +2	Sullivan +2
Arkansas	Cotton +7	Tie	Cotton +9	Cotton +8	Cotton +3	Cotton +6
Colorado	Gardner +2	Gardner +8	Gardner +2	Udall +4	Udall +3	Udall +1
Georgia	Perdue +1	Nunn +3	Nunn +2	Tie	Perdue +6	Perdue +12
Iowa	Ernst +2	Braley +2	Ernst +4	Braley +3	Ernst +2	Tie
Kansas	Orman +1	Roberts +4	Orman +4	Tie	Roberts +1	Orman +1
Kentucky	McConnell +6	McConnell +7	McConnell +5	McConnell +3	McConnell +3	McConnell +5
Louisiana	Cassidy +5	Cassidy +3	Cassidy +9	Landrieu +1	Cassidy +5	Cassidy +7
New Hampshire	Shaheen +2	Shaheen +1	Shaheen +4	Shaheen +4	Shaheen +2	Shaheen +3
North Carolina	Hagan +1	Tie	Hagan +4	Tillis +4	Tillis +6	Hagan +3

Shaded polls diverge from the mean by >3.5 points (one standard deviation)

People would have freaked out if some of these results were released in the end stages of this year's campaign. Look at all those "outliers":

- One poll has Republican Cory Gardner up by 8 points in Colorado. Another one has Democrat Mark Udall ahead by 4!
- Republican Bill Cassidy leads by 9 points in one Louisiana poll. But Democrat Mary Landrieu is leading in another!
- Two polls have the Democrat Michelle Nunn ahead in Georgia. Another has Republican David Perdue up 12 points!

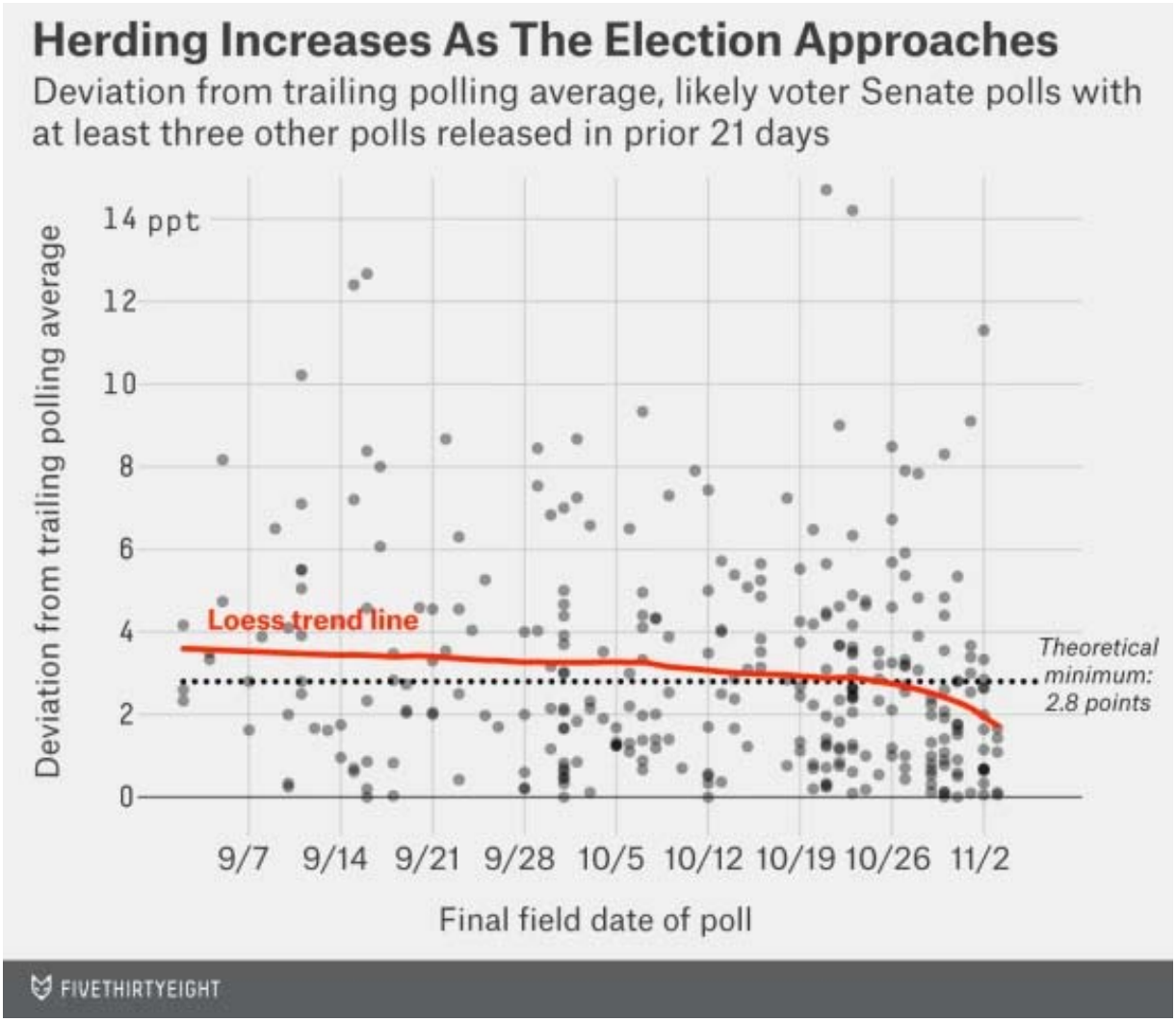
Are these polls skewed? No. This is exactly what polls should look like under ideal conditions — when sampling error is the only thing they have to worry about. Under real-world conditions, where there are other sources of uncertainty, the polls should vary even more than this.

But by the end of the campaign this year, the polls in most states varied only within a narrow range. The next chart describes how much polls across all this year's Senate races deviated from the polling average in their states at the time they were conducted.[6] The polling average is calculated as follows: Each new poll is compared against other polls of the same state conducted somewhere between one day and 21 days beforehand.[7] In calculating the average, I used a maximum of one poll (the most recent one) from each polling firm and didn't compare a polling firm against itself.[8]

These averages should closely resemble those from sites like Real Clear Politics. Note that they are trailing averages — they don't reflect any data from after the poll was

conducted. I'm limiting the analysis to cases where there had been at least three other polls conducted in the previous 21 days, enough that we have a reasonable sense of where the consensus stood on the race.

There's a lot going on in this chart, so let's take a look and then talk about it.



The gray circles represent the results of individual polls (about 300 of them). They show the absolute difference between the new poll and the polling average (without regard to whether the new poll was more Democratic- or Republican-leaning than the average).

The data is noisy. But the red trend line, which is based on loess regression,[9] reflects how much polls were typically deviating from the polling averages. In early September, for example, the typical poll deviated from the polling average by about 3.5 percentage points.

As the election season wore on, new polls hewed somewhat more closely to the polling

averages. But the change was marginal until the final week or two of the campaign, when they started to track it much more closely. By the eve of the election, new polls came within about 1.7 percentage points of the polling average.

Perhaps you could construct some rationale, apart from herding, for why the polls behaved this way. Maybe it became easier to predict who was going to vote and that made methodological differences between polling firms matter less. As a more technical matter, the volume of polling increased as the election approached; this presents some complications, which I address in the footnotes.[10]

But there are two dead giveaways that herding happened. One is the unusual shape of the curve. Rather than abiding by a linear progression, it suddenly veers toward zero in the final week or so of the campaign.

What happened during this period? It's when pollsters were releasing their final polls of the campaign — the ones they think posterity will judge them by. These polls were included in the final Real Clear Politics averages and received a heavy weight in the final FiveThirtyEight forecast.

The impolite way to put it is that this was CYA (cover-your-ass) time for pollsters. Some that had produced "outlier" results before suddenly fell in line with the consensus.

The other giveaway is the one we discovered before in Iowa. By the end of the campaign, new polls diverged from the polling averages by less than they plausibly could if they were taking random samples and not tinkering with them.

As I mentioned before, an 800-person poll has a standard error of 3.5 percentage points because of sampling error alone. A related calculation is the *average error* introduced by sampling. As a rule of thumb, the average error is equal to about 80 percent of the standard error — or in this case, about 2.8 percentage points.

This is the theoretical limit on how accurate polls can be. Even if pollsters knew, for instance, that David Perdue would win by exactly 7.9 percentage points in Georgia (as he did), they'd still miss this result by 2.8 points on average when collecting 800-person samples.

In fact, however, the new polls deviated from the polling average by less than 2 points by the end of the campaign. How did that happen? To be clear, I'm not accusing any

pollsters of faking results. But some of them were probably "putting their thumbs on the scale," manipulating assumptions in their polls such that they more closely matched the consensus.[11]

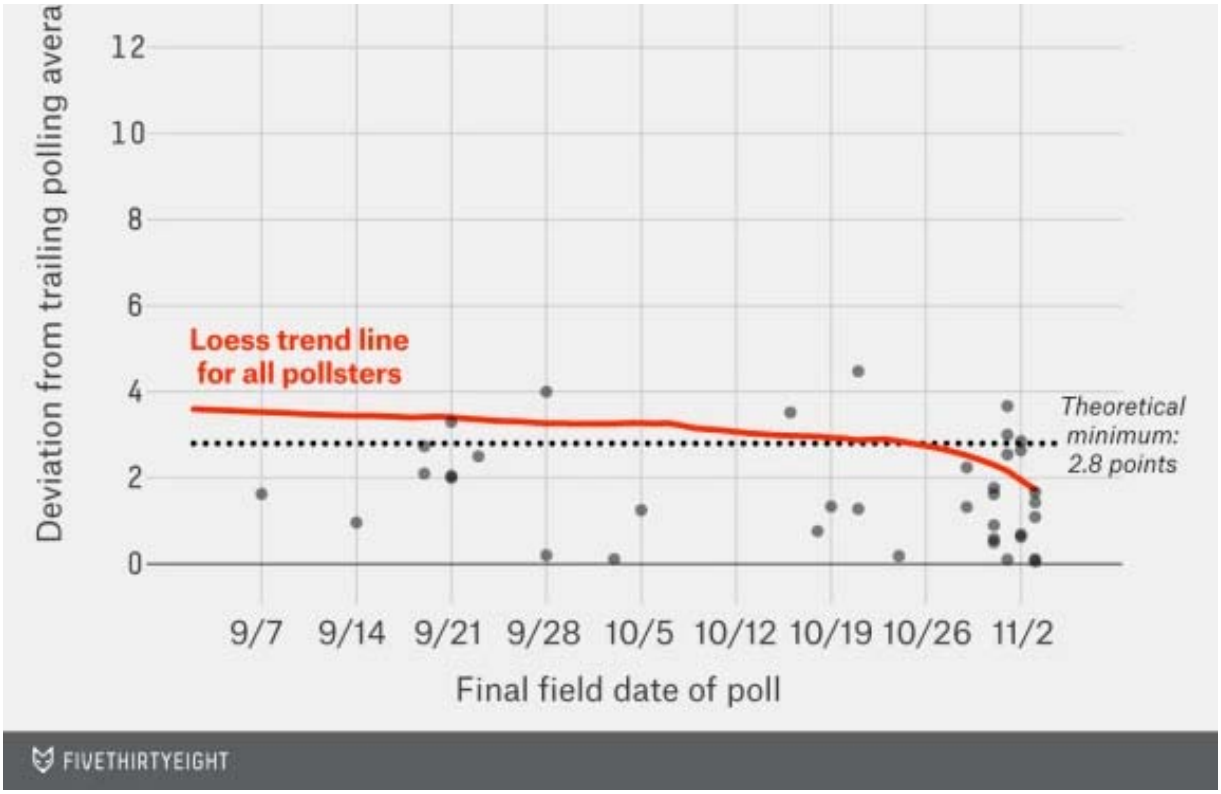
In some cases, the pollsters' intentions may have been earnest enough. Perhaps they ran a poll in Iowa and it came back Ernst +7. *That can't be right*, they'd say to themselves. *No one else has the race like that*. So they'd dig into their crosstabs and find something "wrong." *Ahh — that's the problem, not enough responses from Ames and Iowa City.*[12] *Let's apply some geographic weights. That comes out to ... Ernst +3? We can live with that.*

Even when the pollsters mean well, this attitude runs counter to the objective, scientific nature of polling. As a general principle, you should not change the methodology in the middle of an experiment.

A few pollsters are shameless about their herding. One of them is Public Policy Polling (PPP), a polling firm that conducts automated polls for both public consumption and for liberal and Democratic clients.

Take a look at this exchange, for example, between The New York Times' Nate Cohn[13] and PPP's Tom Jensen. Cohn discovered that in 2012, the racial composition of PPP's polls was correlated in an unusual way with President Obama's performance among white voters in their surveys. If Obama was performing especially poorly among whites in one PPP poll, it tended to have a higher share of nonwhite voters, which boosted Obama's result. And if Obama was doing relatively well among whites, PPP projected less nonwhite turnout, keeping his lead in check. As a result, PPP's polls tended to show an unusually steady race between Obama and Mitt Romney.

I'm picking on PPP for a reason: They're the biggest herders in the business. Here's the chart I showed you before, but with only PPP's polls highlighted. On average, in states with at least three other recent polls, their polls deviated from the polling average by only 1.6 percentage points. The evidence for herding is extremely clear visually and statistically.[14]



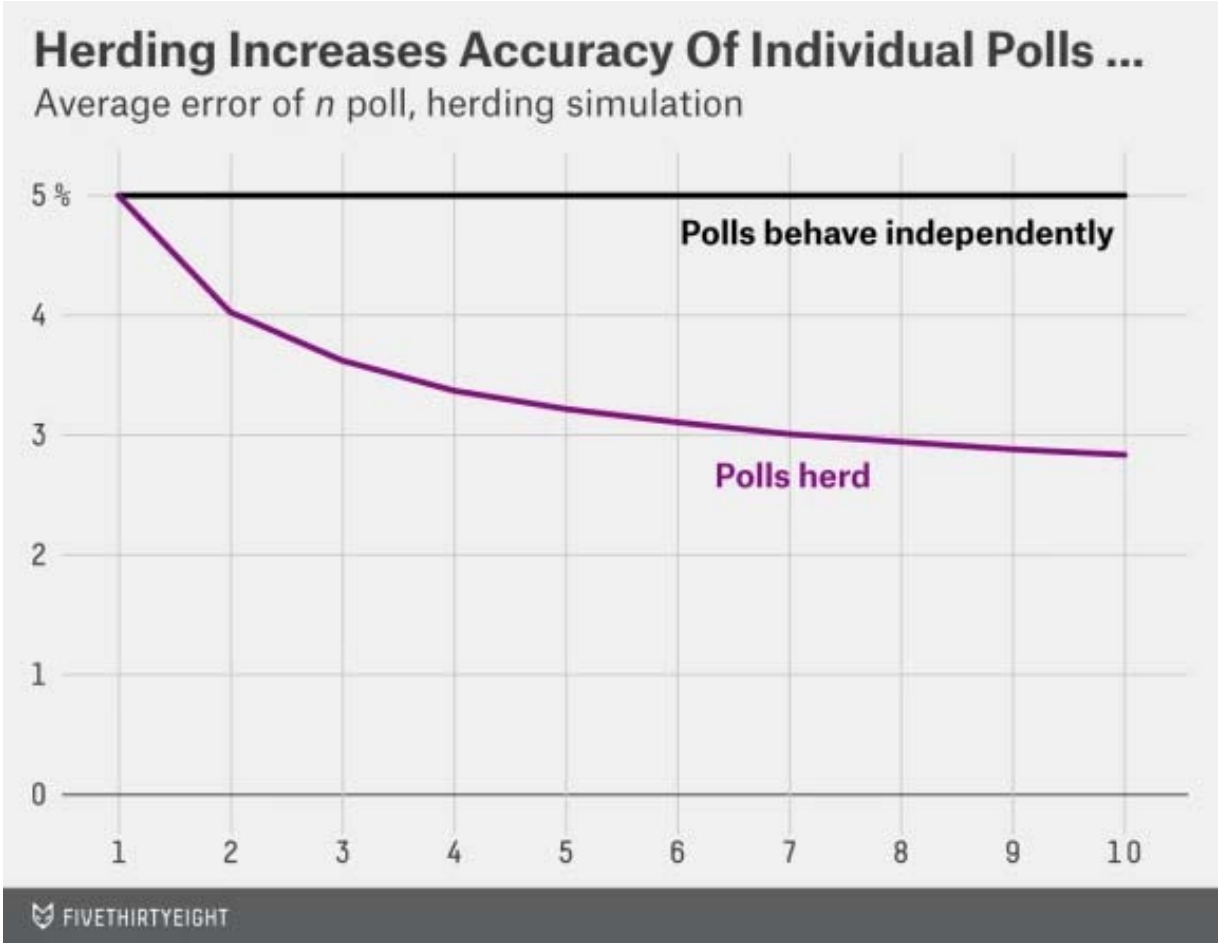
So perhaps Public Policy Polling sits at the opposite end of the spectrum from J. Ann Selzer. But herding may be hard to eradicate. The paradoxical-seeming reason is that herding can make the *average poll* more accurate even as it makes the *polling average* worse. (For economics nerds — this is sort of a tragedy of the commons problem.)

To demonstrate this, I created another simulation in which pollsters engaged in herding and compared it to one where they didn't. The rules of the simulation are as follows:

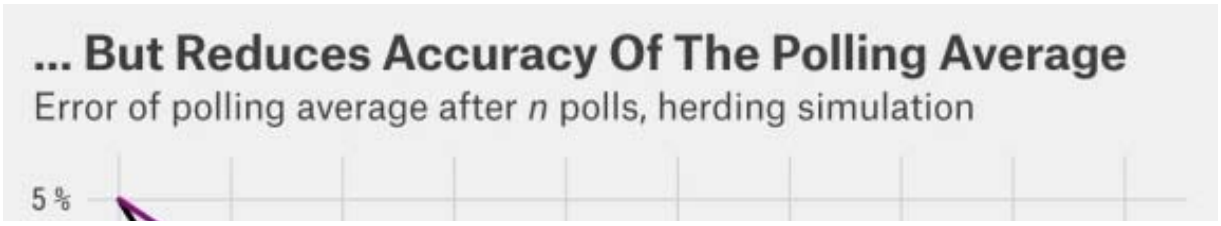
- A series of 10 polls are conducted in sequence (rather than simultaneously) in a state where the Democrat and Republican are tied in the race.[15]
- However, the pollsters don't know the correct result ahead of time. Furthermore, their polls are subject to error. Specifically, their polls miss the correct result by an average of 5 percentage points. This figure corresponds to the historical average error among Senate polls conducted late in the campaign.
- In one version of the simulation, the pollsters behave independently, publishing their number "as is" regardless of what previous polls have said.
- In the other version, the pollsters herd. They do this by truncating the results they publish such that they never deviate by more than 3 percentage points from the

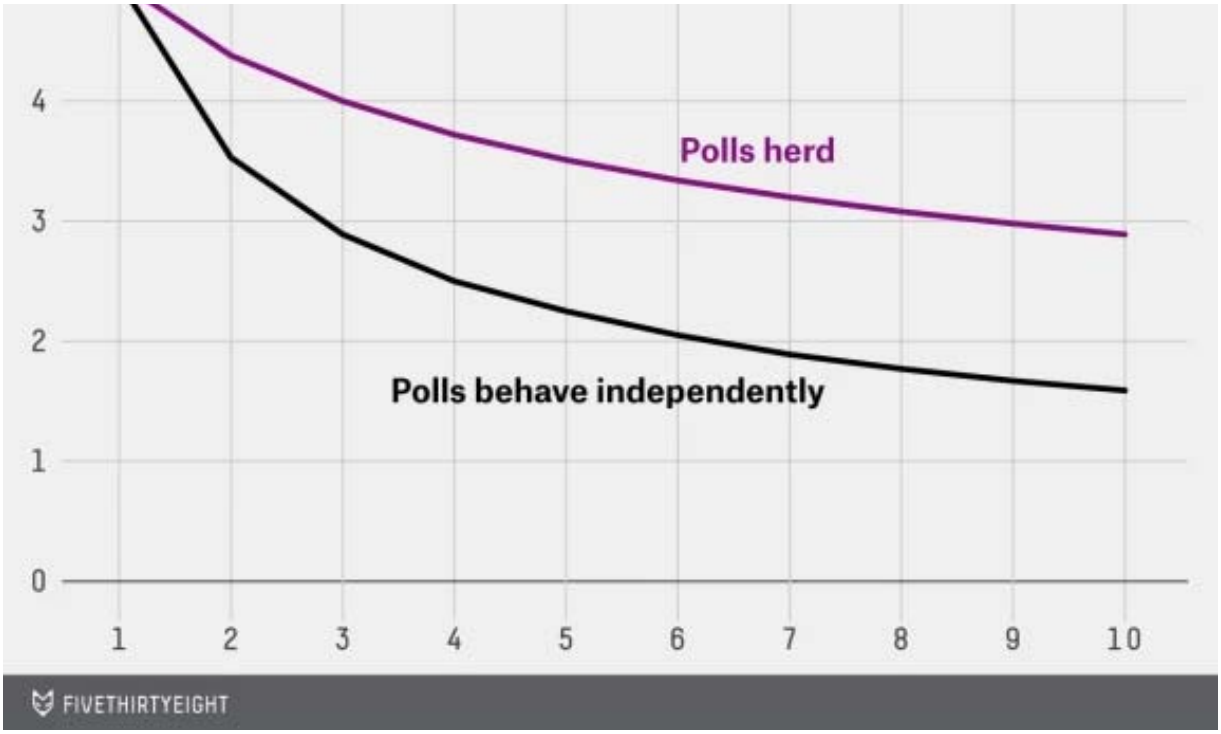
average of previous polls. For instance, if after the first five polls are conducted the Democrat is ahead by 4 points in the polling average, the sixth pollster will never publish a result showing the Democrat ahead by more than 7 points or less than 1 point.[16]

The next chart shows the accuracy of individual polls in the simulation. If the polls act independently, the error stays constant at an average of 5 percentage points for each survey. If they herd instead, the first poll still has a 5-point error (it derives no benefit from herding since there are not yet any other polls to herd toward) but every subsequent poll does a little better. By the time we get to the 10th poll, it misses the actual result by an average of about 3 points rather than 5.



However, as I mentioned, what helps the polls individually hurts them collectively. The next chart shows the accuracy of the *polling average* as opposed to the *average poll*.





In both versions of the simulation, the polling average becomes more accurate as more polls are added to it. But the benefit is much greater when the polls are acting independently. In the independent case, the polling average has an error of just 1.6 percentage points by the time all 10 polls are included in it. If the polls herd, however, the polling average still misses by about 3 points even if all polls are included.

The problem is simple enough to diagnose: When pollsters herd, if the first couple of polls happen to get the outcome wrong, subsequent ones will replicate the mistake. I wonder if these dynamics explain the poor performance of the polls in some states this year. In Kansas, the final polls showed the independent candidate, Greg Orman, ahead in the Senate race by an average of 1 percentage point to 2 percentage points, but the Republican, Pat Roberts, won by 11 percentage points instead.

The first polls conducted of Kansas after the Democratic candidate dropped out of the race were from Public Policy Polling and showed Orman ahead by 10 points. PPP's polls can be highly inaccurate when they don't have other polls to herd toward. In this case, however, other pollsters may have herded toward PPP, producing an incorrect consensus about the race.[17]

This may also be part of why the polls have frequently proved to be "skewed" toward either Democrats or Republicans. In 2012, there was a significant bias toward Republicans and in 2014 a significant one toward Democrats.

This is not a new phenomenon — similar problems occurred in 1980, 1994, 1998 and 2002, among other election cycles. But 2012 and 2014 ought to disabuse us of the notion that the polls are sure to be more accurate just because there are more of them now than there once were. The whole benefit of the "wisdom of crowds" approach depends on people acting independently. A poll that regurgitates polling averages provides no independent information whatsoever.[18]

So ... what to do about it? If you've read this far, you're undoubtedly highly interested in polling. So my message for fellow polling geeks is as follows: Let's not give pollsters so much grief the next time they publish what looks to be an "outlier." Polling data is noisy and polling is becoming more challenging. The occasional or even not-so-occasional result that deviates from the consensus is sometimes a sign the pollster is doing good, honest work and trusting its data. It's the inliers — the polls that always stay implausibly close to the consensus and always conform to the conventional wisdom about a race — that deserve more scrutiny instead.

Footnotes

1. Other researchers have found similar effects. The effect seems to be stronger among polling firms that conduct "robopolls" or use other nonstandard techniques.

2. Calculations like the margin of error, as they usually appear in news accounts, refer to the uncertainty related to one candidate's vote share. For instance, if Ernst is listed as having 55 percent of the vote with a margin of error of 4 percent, it covers a range between 51 and 59 percent. In a two-candidate race like Iowa's Senate contest, however, almost every vote that doesn't go to the Democrat will go the Republican: If a poll misses high on Ernst's vote it will miss low on Braley's, and vice versa. This means the margin of error associated with the difference between the candidates is about twice what's normally listed in the newspaper. In the case of an 800-person poll, for example, the margin of error is roughly 7 percentage points in estimating the margin between the candidates.

3. By this, I mean the result you'd get if you surveyed the whole population rather than taking a random sample.

4. Over the long run, that is. Across a small sample of polls, you could have a quirky result.

5. The 3.5 percentage point figure, to reiterate, assumes a sample size of 800 voters.

6. The analysis is limited to likely voter polls and includes polls like partisan surveys that weren't

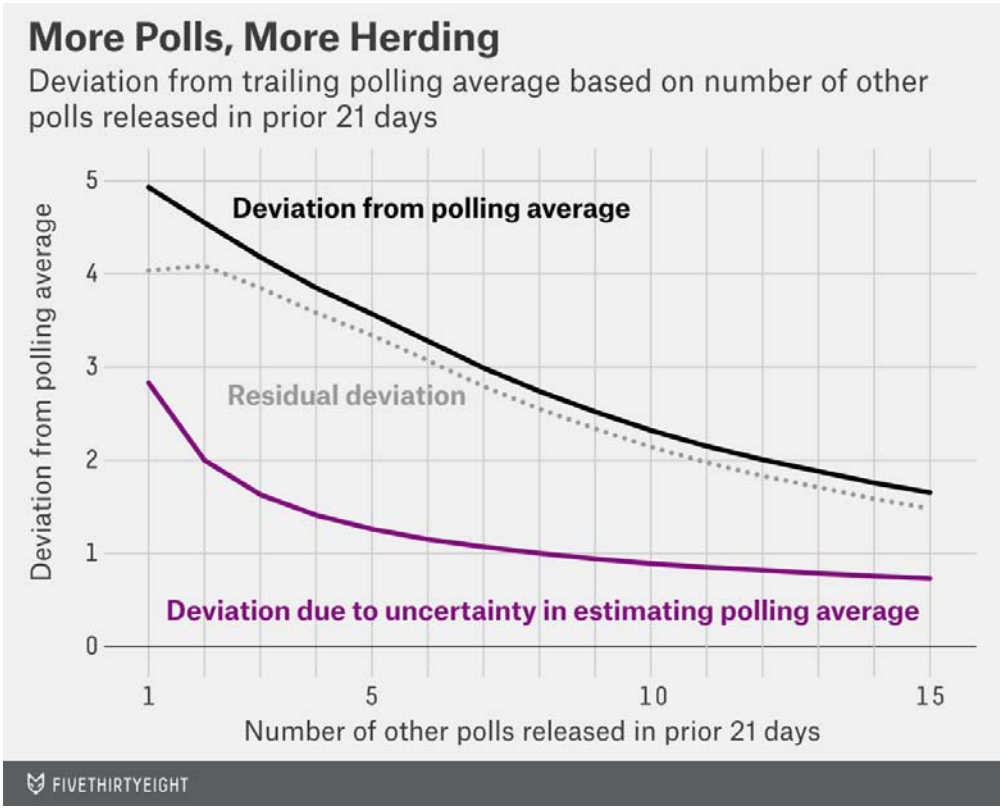
included in the FiveThirtyEight forecasts. I haven't included Rasmussen Reports' controversial Oct. 9 poll of Kansas. For Louisiana, I used polls of the Nov. 4 primary rather than the forthcoming Dec. 6 runoff.

7. Dates are based on the final day of interviewing for the poll, not necessarily the date the poll was released.

8. For instance, Quinnipiac's Oct. 27 Iowa poll was not used in calculating the polling average for its Nov. 2 poll.

9. The loess regression line uses a smoothing parameter of 0.8. This is a conservative setting — the line shouldn't overreact to spurious patterns in the data.

10. In fact, there was a relationship between the volume of polling and the degree of herding. This is depicted in the chart below:



The complication is that the deviation between a new poll and the polling average reflects both the sampling error associated with the new poll and that associated with the polling average. Furthermore, the sampling error associated with the polling average falls over the course of the campaign since more polls are conducted toward the end of the race.

However, herding from new polls increased faster than the improved precision of the polling average did. Assuming the typical poll surveys 800 people, going from three polls to 10 will reduce the standard error associated with the polling average by about 0.5 percentage points. New polls deviated by about 4 points from the polling average when there were three other polls in the field but by just 2 points when there were 10 other polls instead — a considerably

larger decline. The dashed line in the chart reflects the amount of herding that cannot be accounted for due to the increased precision of the polling average, as according to a sum of squares formula.

11. Or they may have been suppressing the publication of polls they perceived to be outliers; see here for what looks to me like a clear example of that.

12. These are college towns where Democrats typically perform well.

13. Cohn was then at The New Republic.

14. In a regression model where the dependent variable is how much a poll deviated from the polling average and the explanatory variable is whether the poll was conducted by PPP, the PPP variable was significant at the 99.9 percent confidence level. (The regression model, like the charts, considered cases where at least three other polls of the state had been conducted during the prior 21 days.)

```
. regress absdeviate pppdummy if other_polls>=3
```

Source	SS	df	MS		
Model	94.606601	1	94.606601	Number of obs	= 326
Residual	2145.77752	324	6.62277013	F(1, 324)	= 14.29
				Prob > F	= 0.0002
				R-squared	= 0.0422
				Adj R-squared	= 0.0393
Total	2240.38412	325	6.89348961	Root MSE	= 2.5735

absdeviate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pppdummy	-1.641937	.4344258	-3.78	0.000	-2.496589	-.7872856
_cons	3.289852	.1521727	21.62	0.000	2.99048	3.589223

Note, however, that PPP polls can deviate wildly from the actual results when they have few other polls to anchor themselves to. In mid-October, PPP published an Idaho poll — just the second poll of the state since Labor Day — showing Republican incumbent Jim Risch ahead by 18 points against Democrat Nels Mitchell. Risch won by nearly 31 points.

15. Note that this assumption makes no difference to our conclusions. We'd come up with the same results if we assumed that the correct outcome was the Democrat winning by 5 points or the Republican winning by 3 points.

16. To be more explicit about the mechanics of this: The simulation assumes that a polling firm takes its initial sample, which is centered on a mean of zero (a tie between the Democrat and Republican) but with a 5-point average error around this mean. (The error is assumed to be normally distributed.) If the initial sample shows the Democrat ahead by more than 7 points, the pollster publishes a result showing the Democrat +7. If it shows the Democrat ahead by less than 1 point (or the Republican ahead by any margin), it publishes a poll showing the Democrat +1. If the Democrat is ahead by somewhere between 1 point and 7 points, it publishes its result "as is".

17. One characteristic of herding is that it will produce fat-tailed errors; usually the polls do well enough but when they miss they can be way, way off, as they were in Kansas.

18. It may not be a coincidence that in 2012, the polls had a Republican skew after Republicans spent the whole year complaining the polls would be biased against them — and that just the opposite happened with Democrats this year. If a pollster is constantly deluged by complaints from one side, it might carefully examine the assumptions it critiques while neglecting others. This year, for example, pollsters may have made sure they were sampling enough minority voters since this was a frequent topic of concern among Democrats. They might not have scrutinized factors, like overly lax likely voter screens, that could have biased their polls toward Democrats.