

# 2016 Presidential Primary Polling Accuracy and Primary House Effect

CHUYUE HUANG University of California, Berkeley

May 8, 2016

## Abstract

*This research project performs a retrospective analysis on polling accuracy of past primaries in the 2016 presidential election and focuses on the impact of key variables relating to survey methodology. This project also analyzes effect of racial composition of different states and partisan affiliation of the pollsters on the "Primary House Effect," which measures the bias in pollsters toward the frontrunners of both races, called "Clinton Bias" and "Trump Bias." Lastly, this project cross-examines the consistency of the FiveThirtyEight pollster rating in 2014 and the performance of the rated pollsters in the 2016 primary polling.*

## I. INTRODUCTION

As technology and data become more and more accessible and efficient, public opinion polling now serves two primary purposes. First, it allows politicians to get fast and immediate feedback on public opinion, which, in theory, should increase democratic accountability. Second, polling during campaign seasons also allow candidates to formulate their strategies and target at particular states or electoral groups.

That being said, the 2016 presidential primary election witnesses a media-sponsored horserace among candidates based on the numerous polls conducted by polling agencies across the country. Major poll aggregators, such as Nate Silver's Five-Thirty-Eight and HuffintonPost's HuffPoster, also become more and more visible in the public. Poll data preceding the primary elections gathers much public attention and can also have an impact on voter preference and thus election outcomes. This phenomenon, commonly in the political science literature, is called "The Bandwagon Effect," which points out that voters who do not pay much attention to politics but keep track of polls are more likely to vote for the candidate that performs the best on the polls, simply because he thinks that voting for the most likely winner validates his vote. However, pre-election polls should aim to predict, rather than influence, the election outcome. The expected value of the polling results should be the final results of the election. Yet, due to the constant changes in public opinion, one would expect that polls that are closer to the election date are able to predict the results better.

But Sanders's major upset in the Michigan primary stuck a harsh blow against polling accuracy. Whereas every single poll conducted in the month before predicted that Clinton would win by more than five percentage point, Sanders ended up winning the primary by 2 percent. As a pseudo-Sanders supporter myself ("pseudo" because I cannot vote as a non-citizen), both the suspense building to the primary and the dramatic excitement after hearing about his upset led me to question the accuracy of the pollsters, especially in this primary season. At the same time, although I consistently perform very poorly on the primary prediction activity for my political science course on public opinion and voting, I have become more and more interested in researching into the methodologies and factors that influence polling accuracy.

Additionally in 2014, Nate Silver and his FiveThirtyEight team published a list of ratings for over 300 polling agencies (Silver 2014). In an article accompanying the ratings, Silver also noticed three changes within the polling industry. First, he notices the case against Internet-polls has grown much weaker. Second, it becomes harder to distinguish a partisan poll. For example, polling firms, like Public Policy Polling, regularly conducts polls on behalf of interest groups and campaign but do pay for others by themselves. Third, Pollster herding becomes a more serious problem: polling firms are more influenced by each other than before. Do the patterns that Silver noticed back in 2014 still persist today in 2016? And how accurately are the polls in predicting the primaries? These are the questions that I seek to address in my paper.

There are three major components of my paper. The first part analyzes the relationship between polling methodology and other features of the poll (independent variables) and polling accuracy (dependent variable), which I will measure by using the absolute error between the poll result and the final result. I will then utilize multivariate linear regression to examine the effect of several independent variables on polling accuracy, including time (calculated as the number of days between the poll and the primary), sample size, modes of poll, and type of target population. In the second part, I will examine if there is a "Primary House Effect" that gives rise to a pollster bias toward Clinton and Trump, the two front-runners of each party. I am interested in knowing if the polls are systematic biased toward Clinton. I will do so by looking how margin between Sanders and Clinton and between Trump and the runner-up in the poll compares to the margin the final results. I will also analyze how the various features of the race, such as state, type of race and the percentage of racial minority, have an impact on such bias. Lastly in the third part, I will reference to the pollster rating by FiveThirtyEight and see if pollsters with a higher rating by FiveThirtyEight can actually predict the primary results more accurately.

Before I go into my analysis, I will briefly go over how I collected my data.

## II. DATA AND METHODS

### 1. Data Collection

For this project, I have three major datasets. First, I obtained a list of final results of the primaries from RealClearPolitics. Next, I get the poll data of over 80 past and future primary races from HuffPollster. This dataset not only includes the poll results, but also the dates when the polls were conducted and the methods of the polls. The methods of the polls refer to how the data was collected, whether through phone interview or internet. Additionally, I obtained the pollster rating dataset from FiveThirtyEight (referred to as "FTE"), which contains variables that denote the feature of the pollsters, such as their partisan affiliation, as well as other numeric variables that Nate Silver used to calculate the grade assigned to each of the pollsters.

I created a data table named "All," by merging all the data from the RealClear Politics and HuffPollster. While merging the data from FTE, I encountered some difficulty with the names of the polling agencies because the same firm has different names in the two datasets. Eventually, I was able to match the names based on the first word and obtain a list of 27 polling agencies that appears in both FTE and HuffPollster dataset (See Appendix B).

### 2. Basic Data Cleaning

Moreover, in order to perform a deeper analysis on the reasons behind polling accuracy and "Clinton Bias," I also collected information about the type of primary that each state holds. I created a single variable named "Primary." Where as a "1" for this variable means that this particular party

in this state holds a primary, whereas a "0" signifies a caucus. Primary and caucus are different on many levels (Gore 2008). Whereas primary is a statewide process through which voters cast ballots for their preferred presidential candidates, caucus is a local meeting that gather voters to voice their support openly for a particular candidate. Because caucus requires the voters to show up at the gathering, it is often harder for the pollsters to predict turn-out in these caucuses.

I also obtained data from the US Census Bureau on the ethnic composition of each state, which includes five new variables: "White," "Black," "Hispanic," "Asian," "American Indian/Alaska Native," and "Two or More Races." I then merge these variables to the main data table.

Additionally, I created a new set of variables to measure polling accuracy. I took the absolute value of the difference between the poll result and the final result for each of the candidates and denoted the variable as "(Candidate Name) AE" (AE stands for Absolute Error). Similarly, "(Candidate Name) E" is the difference between poll result and final result. As mentioned in the introduction, I intend to measure the "House Effect" for Clinton so I created a new variable called "BiasD" (Bias in the Democratic race), which is calculated as the margin between Clinton and Sanders in the poll result minus the margin in the final result.

### III. POLLING METHODOLOGY

The question that I intent to explore in this section is how accurate the pollsters have been doing in this primary season. To measure their accuracy, I create "absolute error" variable by taking the difference between the final result and poll result for each of the candidates. "AE" will be the dependent variable, which reflects the accuracy, and there are four independent variables of interest that might influence the DV: "Days Until Primary," "Sample Size," "Targeted Voter," and "Modes." I use the difference between the date of the poll and the data of the primary to calculate "Days Until Primary." Sample size is the number of observations in the one particular poll, and I will also use a single dummified variable, called "Likely Voter," to denote the type of population that the poll surveys. A "1" means that the poll surveys likely voters, and a "0" means that the poll surveys registered voters. As of the methods of the survey, I put all the pollsters into four different categories. Three methods are related to phone: "Live.Phone," "IVR (Interactive Voice Response)," and "Automated." The fourth method is Internet. In my analysis, I will also group "Live Phone," "IVR," "Automated.Phone," into another variable called "Phone," so that it will be easier for me to compare the accuracy of phone polls and internet polls.

#### 1. Hypothesis

I predict that as the polls get closer to the primary (as the "Days Until Primary" variable gets smaller), their results will become more accurate. Polls with a larger sample size will also likely to yield more accurate result because of the reduced sampling error. Polls surveying likely voters are more likely to produce more accurately results than registered voters, because those who registered are not more likely to vote in any given election. As for the different modes of survey, I suspect that live phone calls will be the most superior method, which can solicit the highest response rate and reduce the sampling errors. Additionally, phone polls should perform better than Internet polls because of a clearer sample generating process.

## 2. Testing

**Table 1:** *Multivariate Regression on Polling Methods*

	Absolute Error	
	AE	
	(1)	(2)
Sample Size	−0.0001*** (0.00002)	
Days Until Primary	0.103*** (0.002)	
Likely Voters	−0.172* (0.093)	
Automated Phone	1.106*** (0.058)	
IVR	1.076*** (0.061)	
Live Phone	0.611*** (0.052)	
Interaction between Sample Size and Days	−0.00000 (0.00000)	
Phone		0.976*** (0.047)
Internet		
Constant	4.114*** (0.087)	4.709*** (0.044)
Observations	144,375	144,375
R <sup>2</sup>	0.022	0.003
Adjusted R <sup>2</sup>	0.022	0.003
Residual Std. Error	5.557 (df = 144367)	5.610 (df = 144373)
F Statistic	455.899*** (df = 7; 144367)	430.252*** (df = 1; 144373)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

The multivariate regression on the method of polls and other relating factors, as shown in Table 1, confirms some of my hypothesis. First, sample size does negatively correlates with the accuracy, meaning that a poll with a larger sample size will likely to yield more accurate result. With respect

to time, the relationship is positive. This is because as opinion changes over time (Green), voters will accumulate more knowledge about the candidates as election comes closer, which allows polls closer to the primary to have a smaller error. I also test the interaction between "Sample Size" and "Days Until Primary" and see that the two do not correlate with each other. It is possible that as pollsters will spend more money to obtain a larger sample, as it gets closer to the primary and thus get a more accurate result.

### 3. Analysis: Likely Voters vs. Registered Voters

While comparing polls that survey likely voters and registered voters, we can see that improvement in the likely voter model has helped increase polling accuracy. In Table 1, we can see that the coefficient for "Likely Voters" is -0.172, which means that surveying likely voters will produce less absolute error than registered voters.

To the public, the difference between likely voters and registered voters might be subtle, but such difference is of great importance to polling accuracy. In order to obtain a group of individuals who will actually go and participate in the primary elections—the "likely voters", pollsters usually ask "screening" questions about past voting (Clement 2014). In 2014, the Pew Research Center publishes a report (Keeter and Igielnik 2014) that test and compare the different kinds of likely voter models. Specifically, the report points out one variable that all pollsters use in their models, which is the measurement of voter intent. In a typical survey, voter intent can be measured by a series of questions that directly asks the participant to report their voting history (in local and national elections), time spent on following elections, and willingness to participate in the upcoming elections. In addition, the Pew report concludes that actually adding past voting record can help improve the likely voter model to the greatest extent. Therefore, we can conclude here polls with likely voters are indeed more accurate than those with registered voters.

### 4. Analysis: Polling Methods

With regards to the different kinds of poll methods, if we compare the three types of phone surveys, we can see that "Live Phone," with the smallest coefficient, is the one that yields the best result. However, another interesting result comes up in the comparison between phone polls and Internet polls. Because both variables are dummified, a "1" for the variable "Phone" will automatically mean a "0" for the variable "Internet." Yet Table 1 shows that "Phone" has a positive coefficient, which means that Internet polls are actually more accurate than the phone polls. This is an interesting phenomenon that I will explore further in the next section.

In terms of the weaker case against Internet polls, my analysis is consistent with Silver's observation in 2014. The coefficient for "Phone" is 0.976 in Table 1, which means that phone polls are actually less accurate than internet polls in predicting the 2016 primary results. Yet we still have reasons to be skeptical of such result. In a report published by the American Association for Public Opinion Research (Baker et al, 2008), it points out that the most obvious source of error of Internet poll is the total survey error because only one third of the US population has access to Internet. Thus, although most of the Internet polls have improved their methodologies and recruit a panel of correspondents whom they can survey on a regular basis, the volunteering participants might still differ from the general public on key demographics. Indeed, if I take a closer look at my HuffPollster dataset, there are only 176 cases of Internet polls, compared to 1414 cases of phone polls.

I decided to take a deeper look at the three main pollsters that conduct Internet polls: CBS, YouGov and SurveyMonkey. In the published online about methodology used to conduct the primary polls, SurveyMonkey states that the self-selected population who take the survey online

are weighted based on the various demographics to reflect the true US population. Additionally, because there is no sampling error on this particular online poll, SurveyMokey created a bootstrap confidence interval instead. In addition, YouGov conducted eleven waves of surveys from September 2015 to April 2016, and the participants include people who opt into the online panel as well as registered voters who are contacted to participate in the polls. Each participant gets assigned a propensity score based on the various demographic features and then post-stratified upon propensity score deciles. Thus, we can see that good methodology matters more for online polls— with the appropriate weighing mechanism and statistical methods, Internet polls can possibly obtain more accurate results. A 2014 study (Wang et al 2014) by researchers from Columbia University shows that with multilevel regression and post-stratification, they were able to predict the Obama's victory in the 2012 presidential election using an non-representative sample on Xbox more accurately than all the other pollsters, including the poll aggregators.

#### IV. PRIMARY HOUSE EFFECT: CLINTON BIAS AND TRUMP BIAS

After evaluating the impact of polling methods on polling accuracy, I now turn to the possibility for polling bias in the primary race by analyzing the "Primary House Effect." Usually in a national Senate race between two candidates, one from each party, the term "House Effect" refers to pollster's bias toward one party because the margin between the two parties in the final results is smaller than the margin in the polls. Following my discussion on Sanders's upset in Michigan and after reading dozens of emails from Sanders's campaign team that speaks the discrimination in the corporate media, I wish to further explore such "candidate bias" in the polls. However, because the primary race takes place within the two parties rather than between, I will therefore evaluate the bias toward the forerunners in each party, i.e. Clinton and Trump.

From the raw poll dataset, I calculated the margin between Clinton and Sanders, and also between Trump and the runner-up (either Cruz or Kasich in the later races) both in the poll results and the final results. Then, I calculated measurement of bias for both Clinton and Trump by subtracting the margin in the final results from the margin in the poll.

##### 1. Hypothesis

My hypothesis is that in the majority of the democratic races, pollsters do demonstrate a bias toward Clinton, which means that the final margin between the two candidates are usually smaller than what the poll data predicts. Additionally, in states that have a higher percentage of racial minorities and hold caucus instead of primaries, such bias tends to be larger.

## 2. Result

**Table 2:** *Multivariate Regression Model On Bias Toward Hilary between Different Racial Minority Groups and Primary Type*

	<i>Dependent variable:</i>	
	Clinton Bias (1)	Trump Bias (2)
Race:White	−0.187*** (0.012)	0.756*** (0.010)
Race:Black	−0.951*** (0.012)	0.768*** (0.010)
Race:Hispanic	−0.349*** (0.013)	0.951*** (0.012)
Primary Type	−10.204*** (0.197)	−26.189*** (0.191)
Pollster:Dem-Affiliated	−9.039*** (0.347)	−2.749*** (0.302)
Pollster:Rep-Affiliated	4.202*** (0.350)	−2.216*** (0.289)
Pollster:Nonpartisan	−5.977*** (0.337)	−3.707*** (0.283)
Constant	45.593*** (1.180)	−41.332*** (1.040)
Observations	138,125	148,125
R <sup>2</sup>	0.351	0.160
Adjusted R <sup>2</sup>	0.351	0.160
Residual Std. Error	11.238 (df = 138117)	9.288 (df = 148117)
F Statistic	10,690.690*** (df = 7; 138117)	4,028.408*** (df = 7; 148117)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

I performed a multivariate analysis on the four different variables, "White," "Black," "Hispanic", and "Primary." For the variables on race, I left out "Asian," "Asian Pacific Island" and "Two or More Races" because these racial groups are less likely to show a group preference toward either Clinton or Sanders. The results of the regression contradicts my original hypothesis.

If we look at the fourth row on the effect of primary type on accuracy, we can see very clearly that pollsters tend to have smaller bias in primaries, possible because the likely voter model tends to be more accurate in primaries.

### 3. Analysis: Racial Composition

Table 2 gives us some interesting insights into the pollster bias. First, if we look at all racial groups, we can actually see that higher racial composition positively correlate with higher bias toward Trump. And if we look at just black voters, we see that in states with a higher black population, pollsters have been fairly accurately in predicting Clinton's victory over Sanders, as indicated by a negative coefficient of -0.951. This is not surprising because we know that Clinton is fairly popular among the black voters. Thus, if these black voters are successfully turned out by the Clinton campaign team, then Clinton is more likely to secure her pre-primary advantage in the polls and win the primary.

### 4. Pollster Affiliation

If we examine all the coefficients, we can see that nonpartisan pollsters show the least bias toward Trump in the Republican race, whereas Democrat-Affiliated pollsters have the smallest bias toward Clinton. Yet, what's intriguing here is that pollsters affiliated with the Republican party show very favorable bias toward Hilary, as indicated by a coefficient of 4.202. One explanation is that compared to Sanders, Clinton positions herself on the center-left on the political spectrum, in contrast with the extreme leftist Sanders. Pollsters affiliated with the Republicans might thus demonstrate a systematic bias toward Clinton.

## V. POLLSTER RATING

In the last section, I performed a retrospective analysis on the performance of the pollsters that appear in both the HuffPollster dataset and the FiveThirtyEight Ratings dataset. The names of the polling agencies are different in the two dataset: for example, "Emerson College" in one dataset and "Emerson" in the other; "Magellan Strategies" in one and "Magellan Polling.LLC" in the other. I spent a lot of effort matching the pollsters by the first names and eventually narrow down to a list of 27 pollsters that received a rating from FiveThirtyEight in 2014 and also polled some primary races in 2016. A complete list of the pollsters can be found in Appendix A. It is worth noticing that Nate Silver assigned each pollster a grade from "A+" to "F" based on a variable called "Historical Advanced Score," which measure how a pollster's average error has compared to other polling firms surveying the same race. This variable also accounts for the type of election polled, number of days until the election, and the poll's sample size. Since the score is regressed toward the mean, a negative score for this variable indicates above average quality. Additionally, Silver's "pollster grade book" also has the "House Effect" variable, which directly measures if the particular pollster has a bias toward the Democrat or the Republican. A lower score for "House Effect" indicates a favorable bias toward the Republicans, and a higher one, the Democrats. Following the analysis on "Primary House Effect" in the last section, I intend to examine if there's a correlation between "House Effect" and "Primary House Effect."

### 1. Hypothesis

Although the final list of rated pollsters that also surveyed the 2016 races, is small, I suspect that pollsters that received a better grade in 2014 will likely be more accurate in 2016. Also, pollsters with a lower "House Effect" score in 2014 will show a greater bias toward Hilary and smaller bias toward Trump.



## 2. Result

Figure 2 in Appendix C (see the end) is a visual representation of the 27 pollsters, their FiveThirtyEight grade in 2014, and their average absolute errors in the 2016 primary polls. We see that there is generally not much consistency in their grade and their accuracy: I performed a regression analysis to test the linear relationship between the "Historical Advanced Score" and "AE" for the 27 pollsters and found that the coefficient is 0.208, which means that among the 27 pollsters, those rated higher by FiveThirtyEight indeed have a smaller absolute error in predicting the primaries. Because the "Historical Advanced Score" for most pollsters fall in the range of -1 to +1 and the difference between different polling agencies are small, we should interpret this result as rather significant. Indeed, pollsters with a higher rating in 2014 are more accurate in predicting 2016 elections.

## VI. CONCLUSION

In conclusion, this project shows me much insight into the polling accuracy of the 2016 presidential primary. The part of the analysis again reinforces the idea that polling methodology matters to a great extent. Specifically, polls closer to the actual primary give more accurate prediction. Live Phone is the best survey method on the phone that yields the smallest absolute error. Although that the regression analysis shows that Internet polls are better than phone polls, I believe that the data I obtained is not sufficient to prove so. With regards to the "Primary House Effect," I found that pollsters will show smaller bias toward Clinton in states with higher percentage of black voters. In the last section, I found that pollsters rated higher by Nate Silver in 2014 did perform slightly above average than the other pollsters in getting more accurate prediction in the 2016 primary. As an idea for further research into this topic, I intent to carry a similar analysis on polling accuracy after the presidential election is over in November. Since it is almost certain that Clinton will face off Trump in the final race, I will like to test all my findings from all the state primaries in this project again to see if they are still valid in the upcoming national presidential election.

## REFERENCES

- [1] Baker. et al (2010).Research Synthesis: AAPOR Report On Online Panels. *Public Opinion Quarterly* 1-71. Retrieved from <http://poq.oxfordjournals.org/content/early/2010/10/19/poq.nfq048.full.pdf+html>
- [2] Clement, S..(2016, Jan 4) Why the 'likely voter' is the holy grail of polling. Retrieved from <https://www.washingtonpost.com/news/the-fix/wp/2016/01/07/why-the-likely-voter-is-the-holy-grail-of-polling/>
- [3] Enten, H.. (2016, May 9). What The Stunning Bernie Sanders Win In Michigan Means.Retrieved from <http://fivethirtyeight.com/features/what-the-stunning-Sanders-sanders-win-in-michigan-means/>
- [4] NBC News (2016, Jan 5) NBC News | SurveyMonkey Weekly Election Tracking Poll Retrieved from <https://www.scribd.com/doc/294643292/NBC-News-SurveyMonkey-Weekly-Election-Tracking-Poll>
- [5] Mertler, C.. (2002). Demonstrating the Potential for Web-based Survey Methodology with A Case Study. *American Secondary Education*, 30(2), 49-61. Retrieved from <http://www.jstor.org/stable/41064450>
- [6] Murray, G. R., Riley, C., and Scime, A.. (2009). Pre-Election Polling: Identifying Likely Voters Using Iterative Expert Data Mining. *The Public Opinion Quarterly*, 73(1):159-171. Retrieved from <http://www.jstor.org/stable/25548068>
- [7] Keeter, S. and Igielnik, r.. (2016, Jan 7) Can Likely Voter Models Be Improved? Retrieved from <http://www.pewresearch.org/2016/01/07/can-likely-voter-models-be-improved/>
- [8] Silver, N.. (2014, Sep 25) How FiveThirtyEight Calculates Pollster Ratings. Retrieved from <http://fivethirtyeight.com/features/how-fivethirtyeight-calculates-pollster-ratings/>
- [9] YouGov. (2015, Apr 17). The Methodology of the 2016 YouGov/CBS News Battleground Tracker. Retrieved from <https://today.yougov.com/news/2015/09/13/methodology-2016-cbs-news-battleground-tracker/>

Appendix

A. APPENDIX

2016/Academics Sp16/Stats 157/Project/Tables and Plots/plot21.png  
Plot 2.1: Absolute Error in Prediction by States

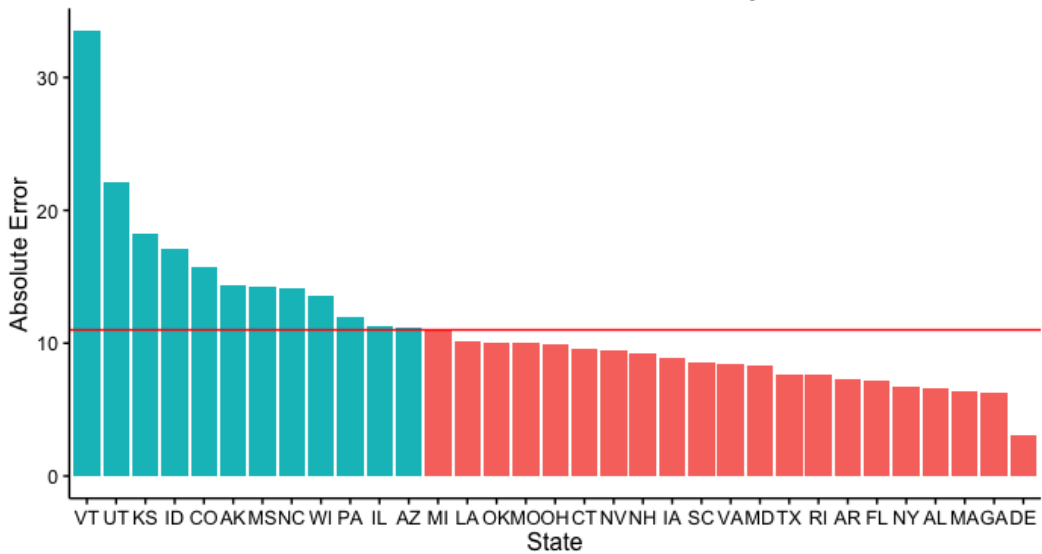


Figure 1: Bar Graph of All FTE Rated Pollsters and Absolute Error

## B. APPENDIX

RatedPollsters	
1	ARG
2	Brown University
3	CBS/YouGov
4	Christopher Newport Univ.
5	Clemson University
6	Dan Jones & Associates
7	Dixie Strategies
8	Elon University
9	Emerson
10	Fox
11	Franklin & Marshall College
12	Gravis Marketing
13	Gravis Marketing
14	Harper Polling
15	Hendrix College
16	Iowa State University
17	Landmark Communications
18	Magellan Strategies
19	Marketing Resource Group (MRG)
20	Marquette University
21	Mitchell Research & Communications
22	Monmouth University
23	Quinnipiac
24	Roanoke College
25	Siena College
26	Suffolk University
27	We Ask America

### C. APPENDIX

2016/Academics Sp16/Stats 157/Project/Tables and Plots/PollsterAE.png

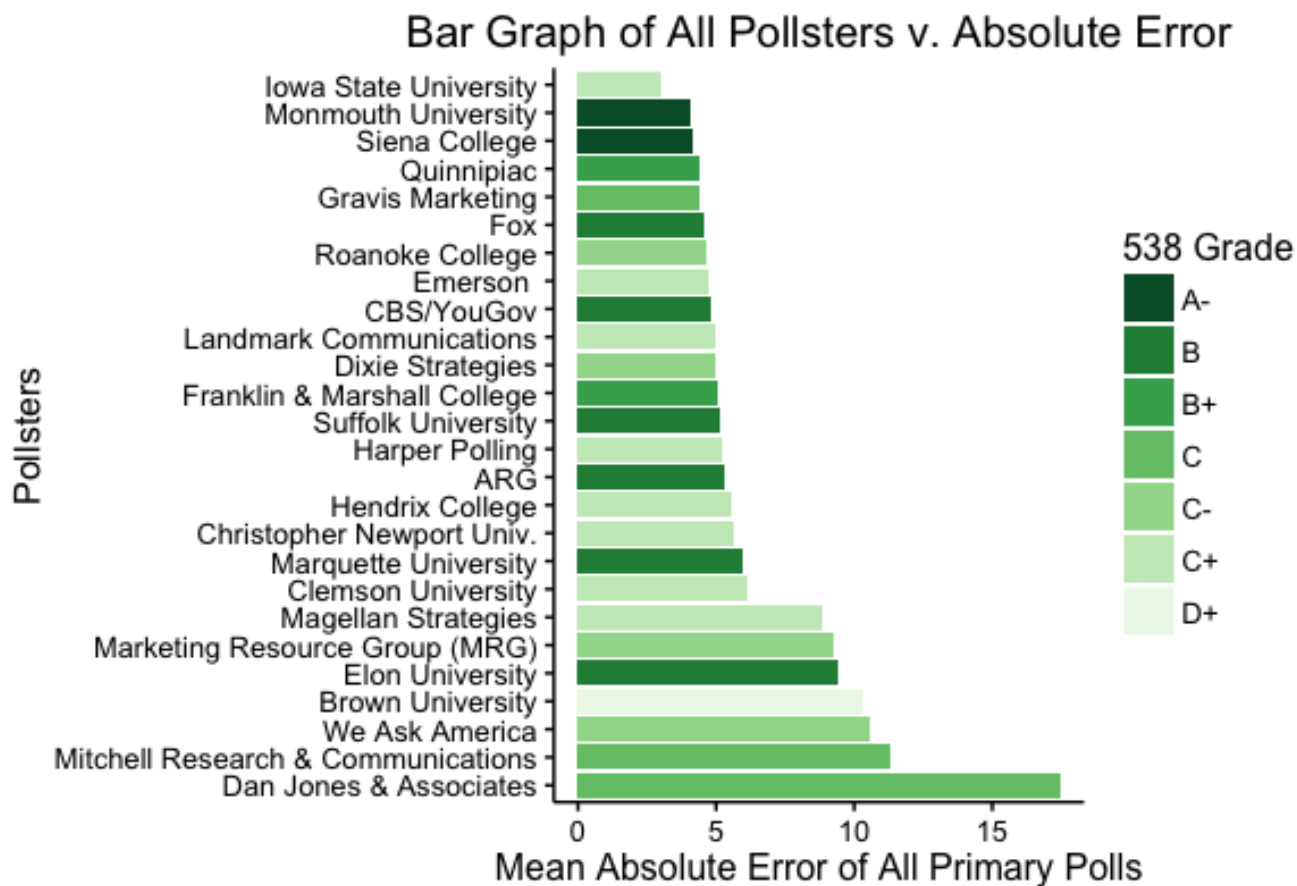


Figure 2: Bar Graph of All FTE Rated Pollsters and Absolute Error