Introduction

Survey sampling is a complex and beautiful art. Comparing the quality of different survey designs is a daunting task, as every design will produce a different estimator. The core purpose of a poll or survey is to estimate the value of the true parameter. Altering the survey design can produce different biases, creating misleading estimators. During the 2016 U.S. Presidential Election, FiveThirtyEight, graded pollsters and adjusted polling results across the U.S., "[using] data and evidence to advance public knowledge, adding certainty where we can and uncertainty where we must," (FiveThirtyEight, n.d.). This report will critique the grading methodology of FiveThirtyEight and its polling adjustments.

Dataset

All the variables from the dataset were used, with the exception of the survey end date, and both raw and adjusted poll scores for Gary Johnson and Evan McMullin. These two candidates were excluded from the analysis due to low candidate support, and in turn their insignificant impact on the outcome of the 2016 U.S. Presidential Election. Survey start date was chosen instead of survey end date to dramatically visualize the number of polls conducted over time (Figure 1a). Additional variables were sourced from other datasets such as the percentage of votes cast for Trump and Clinton from the Federal Elections Commission, or created, such as the margin between Trump and Clinton and true error, which aided in the analysis of FiveThirthyEight's poll adjustments and grading work (Part II).

Polls were not generated simultaneously in the same period. Although this dataset is a time series, considering most of the polling occurred near the election date, we will ignore the time series element of this dataset treating all polls equally. Instead, this analysis seeks to analyze general trends in all polls at an aggregate level; thus, the time series is negligible and would not change the nature of the observations of this report.

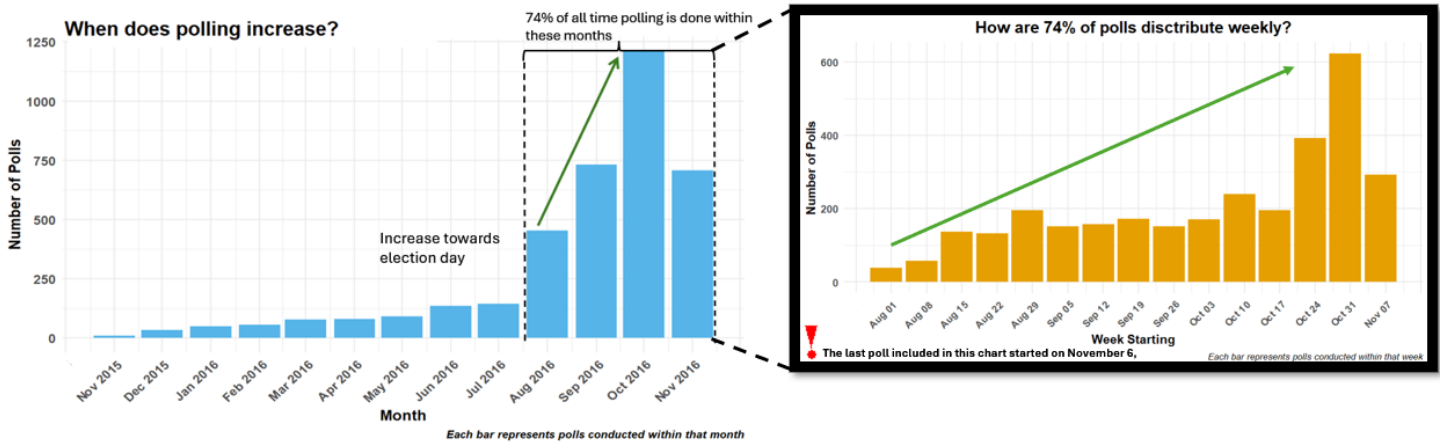
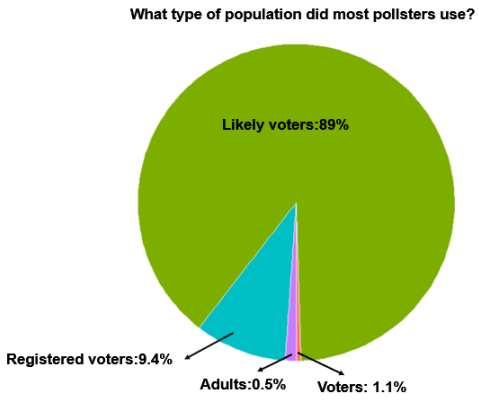


Figure 1a. The number of polls beginning over time monthly, and weekly towards the end of the campaigns.
Figure 1b. The distribution of likely voters, registered voters, voters, and adults by each grade for polls conducted nationwide.

Figure 1a helps users visualize the number of polls conducted over time in the 2016 U.S. presidential election polling, showing a steady increase up until August 2016, closer to the date when Trump was elected as the republican nominee, and towards October 2016, closer to election day. Pollsters separated the surveyed population by classifying them as: likely voters (LV), registered voters (RV), and adults (A) (Kennedy & Deane, 2017). Figure 1b shows the distribution of different population categories, most pollsters used 'likely voters' as their population audience. Predominantly relying on likely voters as their sampling



Note: for the description, definition, and attributes of any data element, refer to the Table 1. Data Dictionary in the Appendix

population, could have led to methodological imperfections, such as the issues discussed above.

Data Analysis

Part I – Overall Prediction Success

Nationally, polls predicted that Clinton would win with a 4.35% lead over Trump, which was adjusted down to 3.2% by FiveThirtyEight. Overall, this wasn't entirely inaccurate, since Clinton did win the popular vote by approximately 2.1%. At the state-level, polling accurately predicted the winner of most states but failed in key 'toss-up' states – notably Wisconsin, Michigan and Pennsylvania. These states all predicted Clinton would win by a comfortable margin going into election day, which surprised many when Trump was eventually declared the winner by a very small margin. Although FiveThirtyEight did adjust raw poll results, it didn't adjust them enough. This was at least in part due to not fully considering how polling might be biased, as well as questionable adjustment methodology (Mercer, 2016).

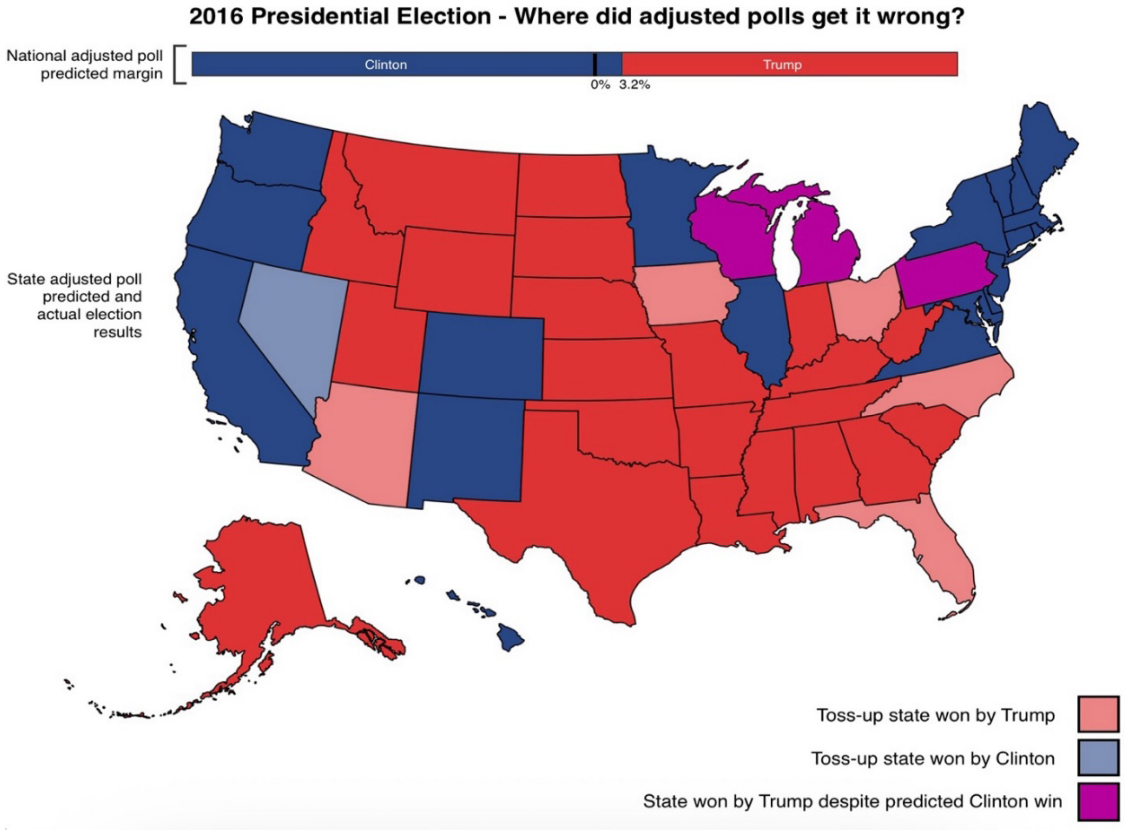


Figure 2. 2016 U.S. presidential election adjusted poll prediction by state.

Part II – Voter populations

The term 'likely voter' is defined by the American Association for Public Opinion Research (AAPOR) by assigning each voter a score based on questionnaire indicating their likelihood to vote (AAPOR, 2022). After making an estimate on the level of turnout, pollsters determine a cutoff point corresponding to the turnout percentage, where the highest scorers are classified as 'likely voters' (AAPOR, 2022). This term has been heavily criticized post 2016 U.S. presidential election polls, as pollsters proposed several theories suggesting it may not have accurately reflected whether those voters would actually vote.

Note: for the description, definition, and attributes of any data element, refer to the Table 1. Data Dictionary in the Appendix

One of the theories was that Clinton support was overestimated in the Midwest (Franck, 2024), most likely due to the assumption that voters with higher education were more likely to support Clinton (AAPOR, 2017). Pollsters overestimated the volume of highly educated voters in the Midwest and underestimated the number of voters with lower levels of education that would come to vote. Another factor at play was that highly educated individuals were more likely to participate in surveys than less educated individuals (Pew Research Center, 2015). Put together, if highly educated people were more likely to support Clinton and more willing to participate in surveys, this would explain some of the overestimated Clinton support in the poll score data. The issue could have been resolved with proper weight adjustments, however, according to AAPOR, the weighting was either improperly adjusted or not adjusted at all (AAPOR, 2017).

Another theory in the discreditation of the term "likely voter" was the "shy Trump voter" theory (Miller & Altman, 2016), which is the under-representation of Trump supporters that may have been caused by the way the polls were conducted (e.g. over the telephone). This theory suggests voters may be more reluctant to show Trump support over the phone due to fear of social judgement. This explanation is consistent with the 'Bradley effect' where certain voters hide their true choices in order to not be viewed as discriminatory, which is evident in the polls on Black candidates' performances that were lower than expected (Miller & Altman, 2016; Altman, 2008). The same can be said for reluctant Trump supporters whom, out of concern for social judgment, may have chosen not to disclose their real views to pollsters.

Part III – Quality grading methodology properties

A sound grading methodology is expected to have the following:

a) Higher (better) grades should have better prediction accuracy
b) Pollster adjustments among higher (better) grades should be smaller
c) There should be a normal distribution of the grades, in this case centered around "B"
d) The error of adjusted poll data should be less than the error of raw poll data
e) Polling in 'toss up' states should show the margin between the main two election contenders being around 0%

a) One would expect polls with higher grades to have higher rate of accuracy, decreasing monotonically as grades decrease. The following shows the probability of each grade successfully predicting which candidate would win the election.

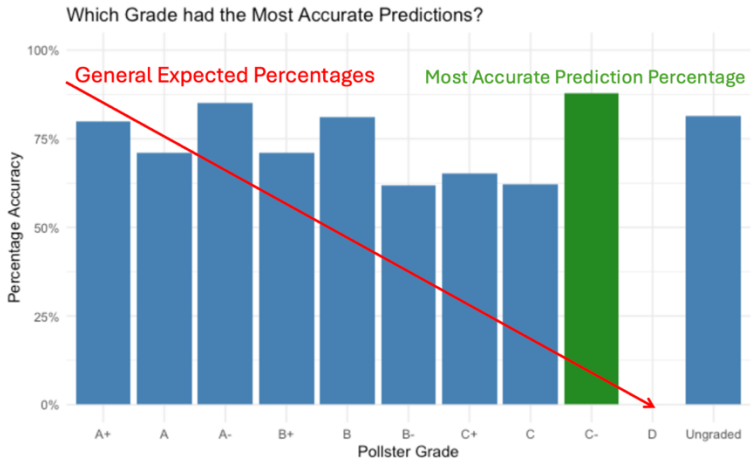


Figure 3. Prediction accuracy percentages for each grade with expected downward trend (red line)

Note: for the description, definition, and attributes of any data element, refer to the Table 1. Data Dictionary in the Appendix

Figure 3 shows that it is not the case, with C- leading all grades in accuracy rates, and A- and B as close seconds. This comes off as especially surprising, because C- the second lowest possible grade in this dataset, showing how misleading grading can be.

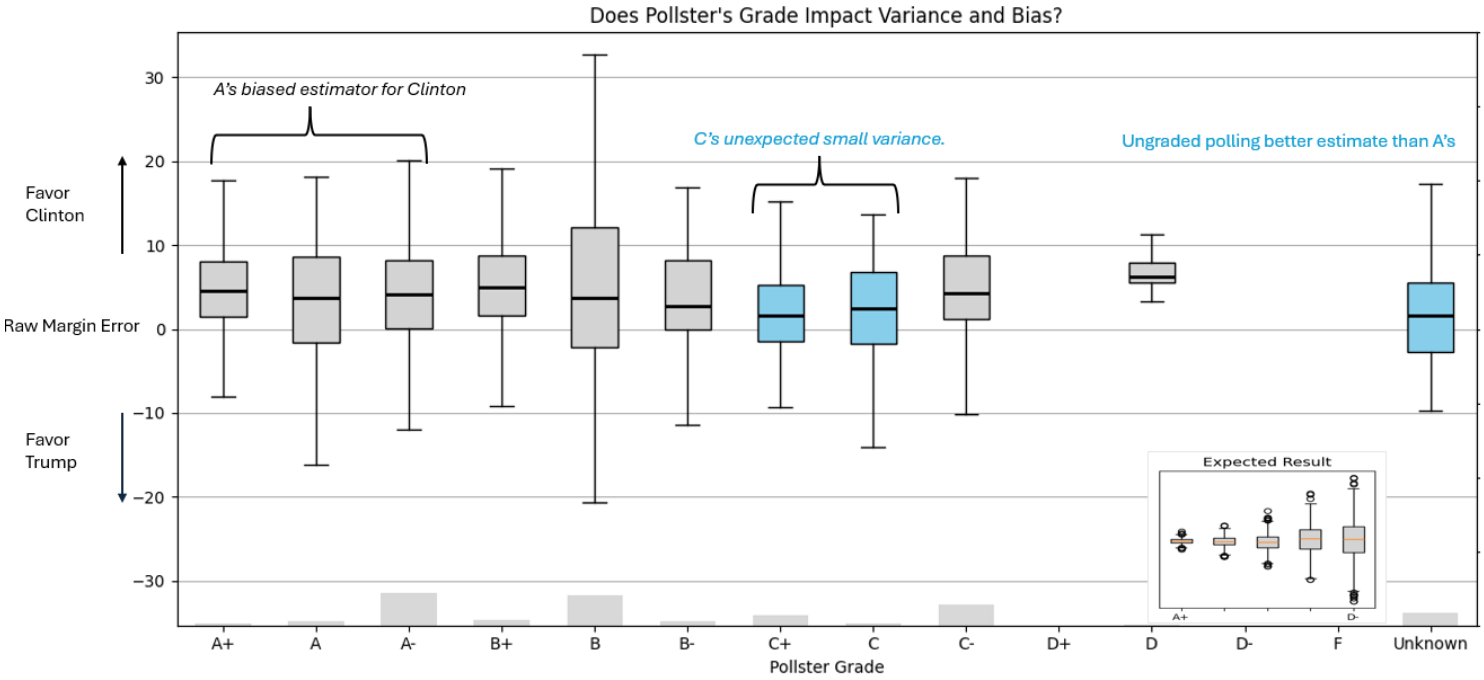


Figure 4. Boxplot of distribution of raw poll errors for each grade

We notice residuals are normally distributed across grades, however, we would expect the variance of those distributions to be smaller for higher grades and increase monotonically with lower grades. This is not the case. Another observation is the C's seem to have less bias – indicating they are a better predictors of margin outcome, and ultimately, election outcome. Again, we would expect A's to be a better-quality estimator meaning less bias and less variance. From here we conclude that C+ and C are in fact the best estimators for margin, not the A grades.

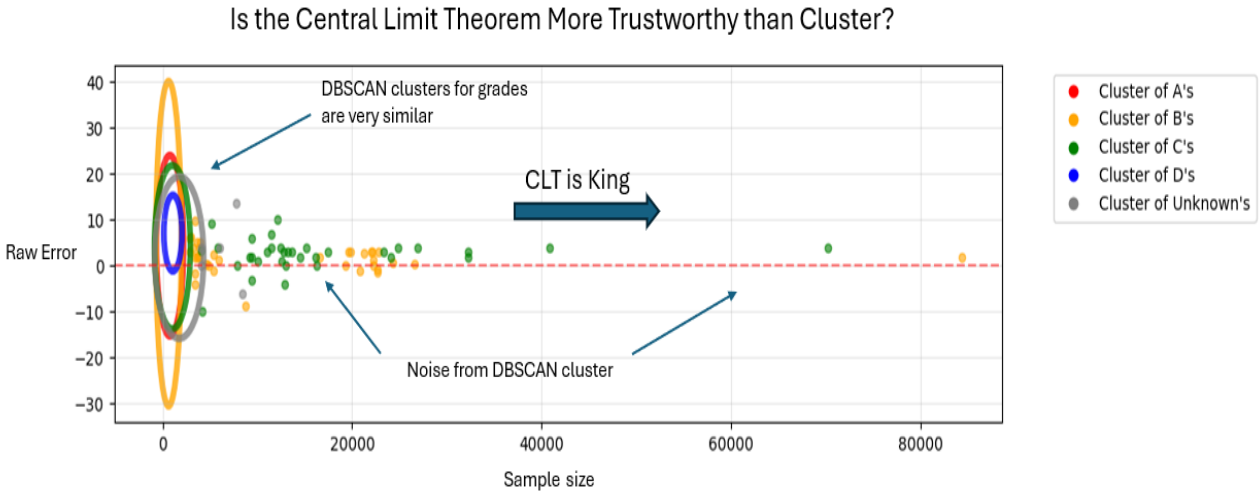


Figure 5. Average Raw Poll Error vs Average sample size

Note: for the description, definition, and attributes of any data element, refer to the Table 1. Data Dictionary in the Appendix

If the grade itself doesn't explain the prediction power of the pollsters, perhaps the central limit can. By clustering polls together in the error-sample size space, we notice there is no real significant relationship between sample size and grade. Thus, we would expect clusters to follow a similar pattern to the contours of a multi-normal distribution in the R^2 plane – A grades should be clustered in the center enclosed by lower grade curves. As expected, the error for polls with very large populations converges to 0. Thus, population size, especially when greater than 10,000 is a better predictor of the quality of the predictive power of a poll than the grading itself. Possibly because the CLT eliminates systematic errors for a sample design.

b) Net margin adjustments for higher grades should be smaller, meaning that higher graded polls are expected to need less adjustment since they are assumed to be "better".

Next, we consider the impact of FiveThirtyEight's adjustment on the predicted margin of victory. Adjustments serve as corrections to potential bias errors in methodology, aiming to produce more accurate estimates of voter intent. For this property, we expect adjustments to be lower for higher graded polls, and vice versa. The boxplot in **Figure 6** shows that there's no monotonic increase in adjustment. This implies although FiveThirtyEight have given pollster B grades, they don't believe they are worse than A's as A are being adjusted more.

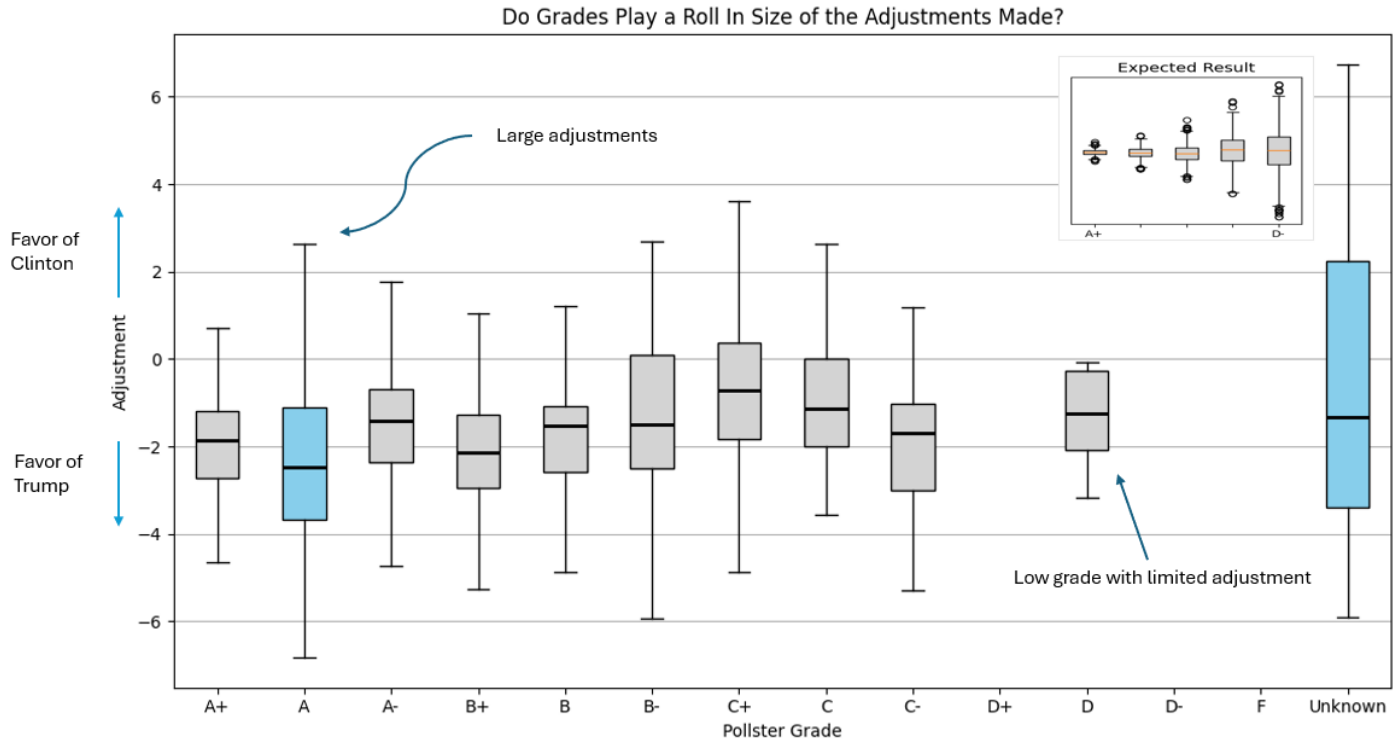


Figure 6. Boxplot of adj. poll margins by pollster grade

c) Like typical grading system we expect a normal distribution of the grades, meaning the assumption is that the expected grade for a given poll is B for the grade range of this dataset.

Note: for the description, definition, and attributes of any data element, refer to the Table 1. Data Dictionary in the Appendix

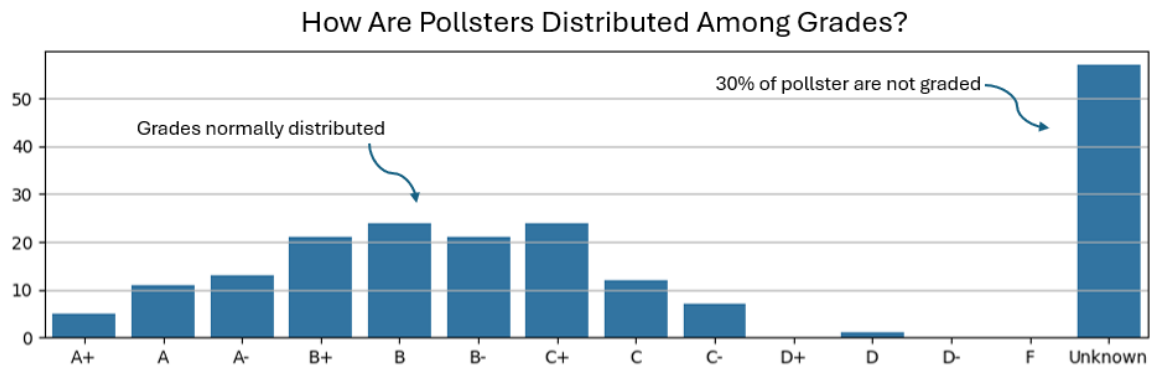


Figure 7. Frequency of assignment of each grade

Figure 7 shows that the dataset follows approximately a normal distribution centered around B/B-. This suggests the criteria used to assign a pollster grade has basis, however that basis does not lie in the pollster ability to reduce bias or variance. Note that the significant portions of pollster were not graded making the overall grading system shaky at best.

d) <u>Error of adjusted poll numbers should be less than Error raw poll numbers.</u>

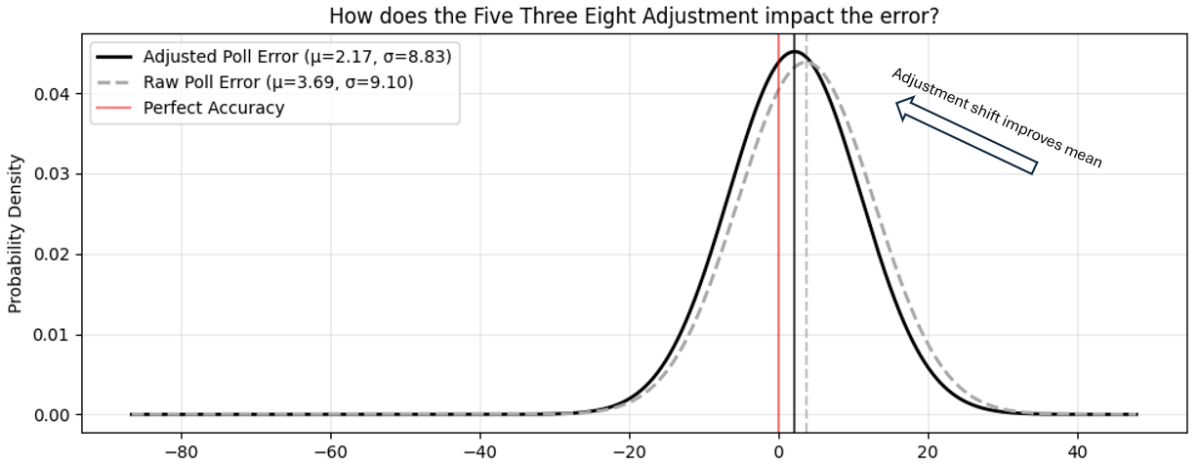


Figure 8. Probability density of (adjusted) poll error

FiveThirtyEight's adjustment successfully reduced the error in the polling estimates, improving the estimators. However, this adjustment was not enough to offset the heavy bias towards Clinton. 2% error is still a hefty advantage (Mercer, 2016).

e) <u>Polling in 'toss up' states should show the margin between the main two election contenders being around 0%</u>

Battleground states in the U.S. election are states where poll margins are considerably small enough that no winner can be predicted to win with any kind of certainty. Most states can reliably predict every election, whereas battleground states have inconsistent results throughout the course of many elections, so it's expected that in these states the margin between the main two election contenders is around 0%. For this grading property, the cluster should ideally be centered at 0, meaning the outcome is uncertain.

Note: for the description, definition, and attributes of any data element, refer to the Table 1. Data Dictionary in the Appendix

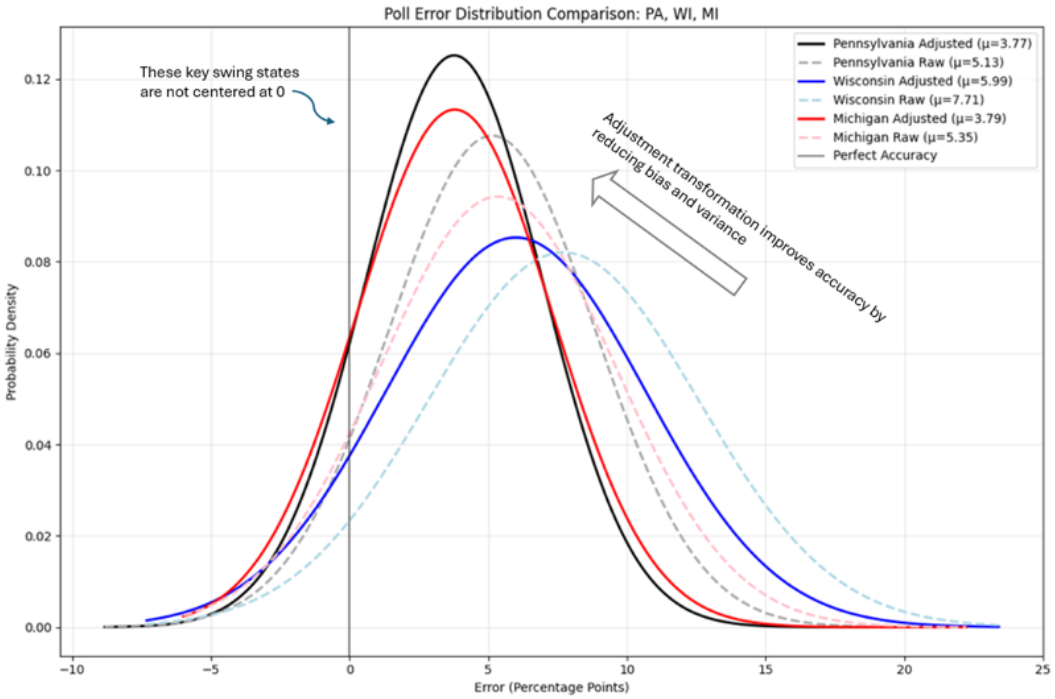


Figure 9. Raw vs. adjusted poll error distributions for Pennsylvania, Wisconsin, and Michigan

Part IV – Super PACs

Super PACs are independent expenditure-only companies/organizations that can raise and spend an unlimited amount of money to support or oppose political candidates. They are unusual in our analysis. They can act either as a useful indicator, or as a sign of where a party is strategically focused. The amount of money a Super PAC raises for a party can suggest how competitive the race is in a state. It can also show where a party and its allies are concentrating their efforts. Based on data from OpenSecrets and the Federal Election Commission (FEC), **Figure 11 shows** the proportions of money spent by campaigns and Super PACs in battleground states.

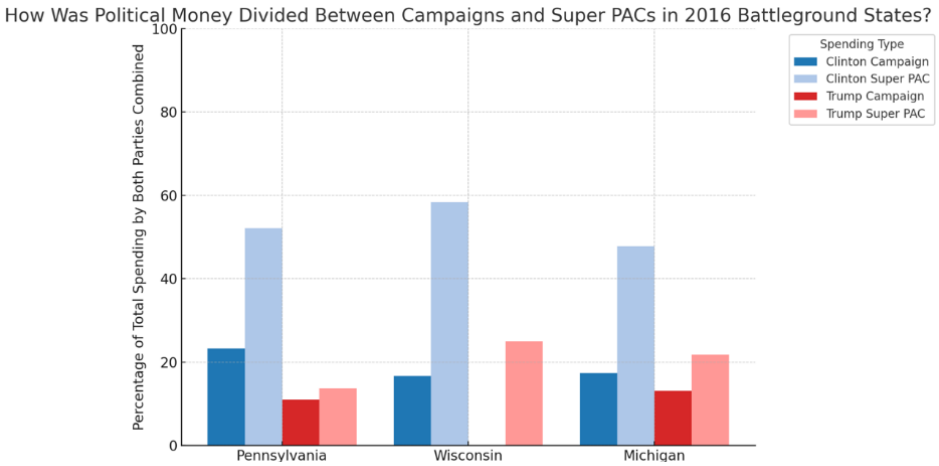


Figure 11: Super Pac vs Campaign Spending for Swing States by Both Trump and Clinton

It becomes obvious that Democratic (Clinton) Super Pac funding was more than that of Republican (Trump) in each of the swing states. This was also the case for campaign spending, as Wisconsin doesn't even have Republican

Note: for the description, definition, and attributes of any data element, refer to the Table 1. Data Dictionary in the Appendix

campaign spending to begin with, showing that they had a different strategic approach, having relied on separate factors on where they needed to invest their resources. Perhaps this implies Democrats believed that these states were closer than the polls suggested.

Conclusion

The analysis done on FiveThirtyEight's adjusted polling data in the 2016 U.S. presidential election cycle shows significant inconsistencies between grading and predictive accuracy. While trying to provide a reliable and objective evaluation of polling quality and account for biases via weighting and adjusting, findings suggest that the grading system provided by FiveThirtyEight is not reliable.

Population terminology was evaluated such as the term 'likely voter', which revealed low prediction power and misconception of voter preferences. With supplemental reading, it was determined that high education and survey participation rates were correlated, and social bias theories could affect survey results, which lead to election predictions that favoured Clinton. Overall, this indicated a larger methodological issue in the likely voter group.

Even though assigned pollster grades closely followed a normal distribution, the performance of higher grades did not uphold the expected results in the following ways. Firstly, pollster grades were not correlated with raw poll errors, or with prediction accuracy. It was expected to observe low poll errors with the best pollster grades, and vice versa, but this was not the case. In fact, grades such as C and C+, demonstrated the least amount of bias and lowest variance for raw margin error. Additionally, boxplots illustrate the adjustments applied to higher-graded polls did not reflect the grade itself as most grades are adjusted similarly. On the contrary, there remained a consistent skew in favor of Clinton, further supporting there may have been a bias in Clinton's favour. Analysis showed sample size as a better indicator for polling accuracy than grades itself. Although the adjustment did reduce the polling error in general, the grading system has too many inconsistencies in its methodology for it to be a reliable indicator of a quality estimator for the general public.

In the age of digital transformation, perhaps polling the way that it has been done is no longer the best estimator. Perhaps stakeholders should look to leverage voter behaviour online as an estimator for election outcome.

Note: for the description, definition, and attributes of any data element, refer to the Table 1. Data Dictionary in the Appendix

Appendix

Table 1. Data dictionary for the 2016 U.S. Election Poll Data Provided to the Team

VARIABLE	TYPE	DESCRIPTION	MINIMUM	MEAN	MAXIMUM	num NAs	% NAs	num DISTINCT VALUES	SOURCE	DATA MANIPULATION
state	character	the area in which a given poll was conducted, by state, state congressional district, or "U.S." for national polls	*na*	*na*	*na*	0	0.0%	57	DATASET	*na*
startdate	date	the start date of the polling period	11/6/2015	8/31/2016	11/6/2016	0	0.0%	352	DATASET	*na*
enddate	date	the end date of the polling period	11/8/2015	9/6/2016	11/7/2016	0	0.0%	345	DATASET	*na*
pollster	character	the organization that conducted the poll	*na*	*na*	*na*	0	0.0%	195	DATASET	*na*
grade	character	the reliability grade given to the pollster based on empirical accuracy (measured by the average error and average bias of a pollster's polls) and methodological transparency	*na*	*na*	*na*	429	10.2%	11	DATASET	*na*
samplesize	number	the number of individuals surveyed in a given poll	35.00	1148.20	84292.00	1	0.0%	1,766	DATASET	Row without a sample size was removed
population	character	the type of population surveyed (a: adults; lv: likely voters; rv: registered voters, v: unknown)	*na*	*na*	*na*	0	0.0%	4	DATASET	*na*
rawpoll_clinton	number	the unadjusted percentage of support for Hillary Clinton	11.04	41.99	88.00	0	0.0%	1,312	DATASET	*na*
rawpoll_trump	number	the unadjusted percentage of support for Donald Trump	4.00	39.83	68.00	0	0.0%	1,385	DATASET	*na*
rawpoll_johnson	number	the unadjusted percentage of support for Gary Johnson	0.00	7.38	25.00	1409	33.5%	584	DATASET	Excluded from analysis, very low support
rawpoll_mcmullin	number	the unadjusted percentage of support for Evan McMullin	9.00	24.00	31.00	4178	99.3%	16	DATASET	Excluded from analysis, very low support
adjpoll_clinton	number	the FiveThirtyEight adjusted percentage of support for Hillary Clinton	17.06	43.32	86.77	0	0.0%	4,199	DATASET	*na*
adjpoll_trump	number	the FiveThirtyEight adjusted percentage of support for Donald Trump	4.37	42.67	72.43	0	0.0%	4,203	DATASET	*na*
adjpoll_johnson	number	the FiveThirtyEight adjusted percentage of support for Gary Johnson	-3.67	4.66	20.37	1409	33.5%	2,209	DATASET	Excluded from analysis, very low support
adjpoll_mcmullin	number	the FiveThirtyEight adjusted percentage of support for Evan McMullin	11.03	24.51	31.57	4178	99.3%	30	DATASET	Excluded from analysis, very low support
grouped_grade	character	the reliability grade given to the pollster based on empirical accuracy (measured by the average error and average bias of a pollster's polls) and methodological transparency	*na*	*na*	*na*	0	0.0%	5	CREATED	*na*
margin_clinton_trump	number	the unadjusted percentage of support for Hillary Clinton minus Donald Trump	-49.95	2.16	84.00	0	0.0%	1690	CREATED	*na*
adj_margin_clinton_trump	number	the FiveThirtyEight adjusted percentage of support for Hillary Clinton minus Donald Trump	-53.02	0.65	82.40	0	0.0%	4203	CREATED	*na*
true_trump_count	number	percentage of votes casted for Trump sourced from Federal Election Commission - left outter join on State	0.04	0.48	0.68	0	0.0%	52	FEDERAL ELECTION COMMISSION	*na*
true_clinton_count	number	percentage of votes casted for Trump sourced from Federal Election Commission - left outter join on State	0.22	0.45	0.91	0	0.0%	52	FEDERAL ELECTION COMMISSION	*na*
true_margin	number	true_clinton_count - true_trump_count	-0.46	-0.04	0.87	0	0.0%	52	CREATED	*na*
adjustement	number	(rawpoll_clinton - adjpoll_clinton) - (rawpoll_trump - adjpoll_trump)	-7.25	-1.51	6.75	0	0.0%	4,208	CREATED	*na*

Note: for the description, definition, and attributes of any data element, refer to the Table 1. Data Dictionary in the Appendix

Table 2. Team Member Contributions to Project 1

Teammate	Contributions
Bilge Kirilmis	Investigation, Visualization, Formal Analysis (population), Writing, Editing
David Liu	Investigation, Visualization, Formal Analysis (super PACs), Writing
Iryna Tkachenko-Riek	Investigation, Visualization, Data Dictionary, Project Administration, Writing, Editing
Liam Dubé	Investigation, Visualization, Formal Analysis (data methodology), Writing

Resources

AD HOC COMMITTEE ON 2016 ELECTION POLLING (n.d.). *An evaluation of 2016 election polls in the United States*. https://aapor.org/wp-content/uploads/2022/11/AAPOR-2016-Election-Polling-Report.pdf

Altman, A. (2008, October 17). *The Bradley Effect*. Time. https://time.com/archive/6686717/the-bradley-effect/

American Association for Public Opinion Research. (2022). *Identifying likely voters in pre-election surveys*. https://aapor.org/wp-content/uploads/2022/12/Likely-Voters-508.pdf

Center for Responsive Politics. (2017, November 27). *2016 presidential race*. OpenSecrets. Retrieved from https://www.opensecrets.org/pres16

Center for Responsive Politics. (2017). *Single-candidate super PACs: 2016 election cycle*. OpenSecrets. https://www.opensecrets.org/outside-spending/single_candidate_super_pacs/2016?chart=V&disp=O&type=C

Desilver, D., & Keeter, S. (2015, July 21). *The challenges of polling when fewer people are available to be polled*. Pew Research Center. https://www.pewresearch.org/short-reads/2015/07/21/the-challenges-of-polling-when-fewer-people-are-available-to-be-polled/

Federal Election Commission. (2017, December). *Federal elections 2016: Election results for the U.S. President, the U.S. Senate and the U.S. House of Representatives*. https://www.fec.gov/resources/cms-content/documents/federalelections2016.pdf

FiveThirtyEight. (n.d.). *About us*. FiveThirtyEight. https://fivethirtyeight.com/about-us/

Franck, T. (2024, May 4). *Why election polls were wrong in 2016 and 2020—and what's changing*. CNBC. https://www.cnbc.com/2024/05/04/why-election-polls-were-wrong-in-2016-and-2020-and-whats-changing.html

Gore, D. (2016, December 23). *Presidents Winning Without Popular Vote.* http://www.factcheck.org/2008/03/presidents-winning-without-popular-vote/

Kennedy, C., & Deane, C. (2017, February 16). *A basic question when reading a poll: Does it include or exclude nonvoters?* Pew Research Center. https://www.pewresearch.org/short-reads/2017/02/16/does-poll-include-or-exclude-nonvoters/

Mercer, A. (2016, September 8) *5 key things to know about the margin of error in election polls.* Pew Research Center. https://www.pewresearch.org/short-reads/2016/09/08/understanding-the-margin-of-error-in-election-polls/

Miller, Z., & Altman, A. (2016, October 25). *Why Donald Trump's latest theories on the election don't hold up*. Time. https://time.com/4545399/donald-trump-election-voting-theories/

Siegel, E. (2016, November 9). *The Science Of Error: How Polling Botched The 2016 Election*. Forbes magazine. https://www.forbes.com/sites/startswithabang/2016/11/09/the-science-of-error-how-polling-botched-the-2016-election/

Silver, N. (2014, September 25). *How FiveThirtyEight calculates pollster ratings*. FiveThirtyEight. https://fivethirtyeight.com/features/how-fivethirtyeight-calculates-pollster-ratings/

Note: for the description, definition, and attributes of any data element, refer to the Table 1. Data Dictionary in the Appendix