THE ELUSIVE ‘LIKELY VOTER’: IMPROVING PREDICTION OF WHO WILL VOTE

An Honors Thesis

Presented by

Anthony Rentsch

Completion Date:
May, 2018

Approved By:

____
Brian Schaffner, Political Science

____
Justin Gross, Political Science

ABSTRACT

Title: **The Elusive 'Likely Voter': Improving Prediction of Who Will Vote**
Author: **Anthony Rentsch**
Thesis/Project Type: **Independent Honors Thesis**
Approved By: **Brian Schaffner, Political Science**
Approved By: **Justin Gross, Political Science**

Political commentators have offered evidence that the "polling misses" of 2016 were caused by a number of factors. This project focuses on one explanation, that likely voter models – tools used by pre-election pollsters to predict which survey respondents are most likely to make up the electorate and, thus, whose responses should be used to calculate election predictions – were flawed. While models employed by different pollsters vary widely, it is difficult to systematically study them because they are often considered part of pollsters' methodological black box. Instead, I build what likely voters should look like from the ground up. Using Cooperative Congressional Election Study (CCES) surveys from presidential and mid-term election years over the last decade I develop a series of likely voter models, and recommendations on how to use them, that allow any pollster to accurately and effectively communicate election predictions by incorporating a range of different turnout estimates.

Table of Contents

1. Introduction

Many political commentators have referred to the “polling misses” of 2016: polls incorrectly predicted the winner in several key swing states, which led some polling aggregators to predict that Democratic candidate Hillary Clinton was a shoo-in for the presidency. While Clinton carried the national popular vote, Republican Donald Trump ended up with the Electoral College advantage and won the presidency. Even if the polls “missed” in 2016, election polling is still a valuable pursuit. Polls provide a snapshot of public opinion during an election cycle, and this snapshot provides valuable data for voters, pundits, and prospective and current elected officials. Additionally, polls are a prime data source for studying campaign decisions and effects.

In the wake of the 2016 general election cycle, political scientists have attempted to assess what went wrong with the polls. Many analyses have been limited to focusing on observable characteristics of these polls, including how many undecided voters polls reported, how the survey was administered, and what demographic variables pollsters used for weighting. Yet, no clear explanation has been identified through this work.

One methodological aspect of election polling that is often discussed in the wake of an election cycle is how pollsters define likely voters. Most polls today report their trial heat estimates in terms of likely voters – among all national and battleground state polls taken during the three weeks leading up to the 2016 presidential election and archived on Huffington Post’s Pollster, 221 out of 227 employed a likely voter model to report their election predictions (Rentsch and Schaffner 2017) – so the choice of how to determine the likelihood that a respondent will end up voting is an extremely important choice. Forecasting who will vote is a particularly challenging problem because, unlike other estimation efforts, pollsters are asked to make an inference about a population (voters) that does not exist yet. Unfortunately, it is difficult to systematically study likely voter models, as they are often proprietary. Even when it is possible to discern what variables a specific poll uses to determine which respondents are likely voters it is even more challenging to discern how they go about doing this. Does a pollster simply treat respondents who report that they “will definitely vote” as likely voters? Do they combine a series of questions to create a composite vote likelihood score and then use the responses that correspond to scores over a certain threshold in their predictions? Are historical trends used to calculate a vote propensity score for each respondent that is used as an additional survey weight? The possible implementations of likely voter models are numerous and this adds to the difficulty of studying the effects of these models on election polling.

With this in mind, I develop a framework for identifying likely voters that builds on political science scholarship and is reasonably straightforward for pollsters to implement in future elections. Motivation for my analysis, as well as for the methods that I use, comes not only from past work that has looked directly at likely voter models, but also at research on election polling, the nature of misreporting voting behavior on surveys, voter turnout, and structural election forecasting.

I leverage Cooperative Congressional Election Study (CCES) surveys taken during presidential and midterm election years over the last decade to build and evaluate different likely voter models. The CCES is a large-N (more than 30,000 respondents each year) nationally representative survey about political attitudes and behaviors, taken in two waves during election years – right before and right after the November election. Fortunately, the CCES asks the standard demographic and vote likelihood questions that election polls typically use in their likely voter models. Additionally, since 2008, the CCES has used vote validation, which is the gold standard for studying election turnout at the individual level. Thus, the CCES is an ideal data set to create likely voter models for both national polls and state-level polls and to analyze the performance of these models under different turnout scenarios.

The final product of my analysis is a series of theoretically- and empirically-backed recommendations for creating likely voter models for election polling at the national and state level. This project should provide pollsters with valuable information regarding which approaches to defining likely voters are more accurate. Furthermore, thinking critically about likely voter methodology can improve how pollsters communicate their results and can help them frame their trial heat numbers in terms of a range of likely electoral outcomes, which is beneficial to how the media and public discourse handle election polls. This project takes seriously calls that polling needs to do a better job of communicating probabilistically, looking specifically at how likely voter models operate under different turnout scenarios (Silver 2017).

On the simpler side, I find that defining a likely voter threshold using a vote likelihood composite index (such as the Pew Research Center's Perry-Gallup index) produces fairly low-error estimates for the national race as well as for state races. On the more sophisticated end, using historical voting data to create a decision tree-based model that outputs a vote propensity score for each respondent, and then using these scores to weight the preferences of all respondents, produces highly accurate national-level and state-level estimates. The vote likelihood composite index approach should be accessible for all pollsters as it takes advantage of questions that are typically asked on election polls for cross tabulations. Additionally, the variables and the approach are easy to explain to a wide audience. The decision tree approach takes a bit more data and is a bit more challenging to explain, but its benefits are such that a pollster should consider investing the resources to implement it. I also show that likely voter models can be easily and attractively used to produce probabilistic predictions of a candidate's margin of victory based on different turnout rates, and that these predictions can be a valuable resource for consumers of polls.

2. Misreporting voting behavior

Nearly all pre-election polls contain a variation of the question “Do you plan on voting in the [name of upcoming election]?” The most simplistic likely voter model would take all of the respondents who indicate that they do plan on voting in the upcoming election and count their responses toward the election prediction. In fact, some pollsters’ likely voter models are as simple as this. Many survey respondents, however, misreport their true voting intentions, making this survey item alone insufficient to determine who will actually vote.

There are a few theories on why misreporting occurs that apply nicely to the context pre-election polls, which are summarized by Ansolabehere and Hersh (2012). When a respondent indicates that they plan on voting, it is possible that they are not lying and that some extraneous event interferes with their prediction of their behavior (Ansolabehere and Hersh 2012). It is also possible that misreporting is primarily attributable to errors that occur during the vote validation matching process, an enormous undertaking in which survey respondents are cross-referenced with a voter file. Government voting records must be thoroughly cleaned, de-duplicated, and some missing data must be input from commercial sources or otherwise imputed before a list of survey respondents can be matched against the voting records. Berent, Krosnick, and Lupia argue that this matching process is error-prone and that self-reported vote behavior is just as accurate as vote validation (2011). Neither of these explanations is satisfying, however. It is hard to believe that enough extraneous events occur to deter a meaningful number of respondents from voting and a validation error-based explanation would not capture why certain groups are more likely than others to have inconsistencies between their reported voting intention and their actual voting behavior (Ansolabehere and Hersh 2012).

The hypothesis that has the most support is that misreporting stems from social desirability bias: respondents answer questions in a way that would be seen as favorable, either to the interviewer or to the respondent. In particular, the people most likely to overreport voting (to say they will vote when they will not, in fact, vote) are "those who are under the most pressure to vote" (Bernstein et al. 2001). Looking at data from the American National Election Study from 1980 to 1988, Bernstein et al. found that this group includes the "more educated, the more partisan, the more religious, and those who have been contacted… to vote" as well as whites in the Deep South, and white Anglos and the relevant minority group who live in areas with high concentrations of either African-Americans or Latinos (2001).

This is a finding that has been replicated frequently, with a few additions to what groups are likely to overreport. Ansolabehere and Hersh (2012) find that "well-educated, high-income partisans who are engaged in public affairs, attend church regularly, and have lived in the community for a while" retrospectively misreport their voting behavior. Other studies suggest that young, black, and nonpartisan respondents tend to misreport (Rogers and Aida 2014), while individuals from older age brackets and who are highly educated are more likely to accurately assess their voting likelihood (Pew Research Center 2000).

Misreporting poses a large problem for pollsters because it is not random error – people who vote more often are demographically and politically different than people who vote less frequently (Pew Research Center 2016) even while there are not substantial differences between self-reported voters and self-reported nonvoters (Rogers and Aida 2014). Since many of the variables tied to misreporting are also correlated with an individual's partisanship, polls that do not report their results in terms of likely voters can often overestimate support for Democrats (Newport 2000). Thus, using self-reported voting intention can produce biased estimates, which is a problem when one out of every four nonvoters reports that they will vote (Freedman and Goldstein 1996).

A lesser concern, but a concern nonetheless for pollsters is also people misreporting that they will not vote. Using data from the 2008 general election, the 2009 New Jersey general election, and 2011 Wisconsin recall election, Rogers and Aida (2014) find that a substantial number of respondents who report that they are unlikely to vote ended up going to the polls. For instance, in 2008 two-thirds of those who self-reported that their likelihood to vote was "50-50" voted, over half of those who self-reported that they would not vote did vote, and over three-quarters of respondents who claimed that they did not know if they would vote casted ballots (Rogers and Aida 2014). Table 1 shows the breakdown of self-reported vote intention versus validated vote for the 2016 CCES. Again, while respondents misreporting that they will vote is a larger issue, pollsters must still consider how to handle those who report that they are less likely to vote.

[TABLE 1 ABOUT HERE]

The issue is one of data quality – a respondent's self-report of their future behavior does not give us reliably accurate information. To compensate for this it is necessary to turn to research on who actually votes.

3. Voter turnout

Literature that focuses solely on the determinants of voting behavior does not map directly onto the pollster's task of separating voters from nonvoters (Crespi 1988). For one, much of this research is focused on factors that are at worst irrelevant to predicting the propensity to vote at an individual level and at best challenging to integrate into a likely voter model. In a review of political science literature on turnout, Blais (2006) notes that political scientists have spent more time looking at the institutional factors (such as parties, compulsory voting, electoral systems, unicameralism, and voting laws) that contribute to turnout rather than the socioeconomic factors. Yet, there are some political scientists who have looked at the role that socioeconomic factors play in voting, and this branch of research is of particular use for pollsters because understanding why individuals vote in the first place is a good starting point for modeling which respondents will end up voting. Literature on how socioeconomic factors are tied to turnout, then, should provide motivation for what variables to consider in a likely voter model.

Canonical works regarding voter turnout have discussed the primacy of socioeconomic status (education in particular) and age in differentiating voters from nonvoters (Verba and Nie 1972; Wolfinger and Rosenstone 1980; Blais 2006). Political activism, ideological extremism, and race are all tied to likelihood to vote, as is the population density in an individual's area (Verba and Nie 1972). Furthermore, marital status and even an individual's occupation can have an effect on the likelihood of voting (Wolfinger and Rosenstone 1980).

Much of this research, however, is outdated and suffers from poor data quality. For one, many of these studies do not use vote validation; they rely on self-reports and are thus subject to the biases that self-reported measures of voting behavior introduce. More recently, Ansolabehere and Hersh (2012) conducted the first ever fifty-state vote validation (they matched the 2008 CCES into the voter file maintained by the firm Catalist, LLC) to revisit the question of who actually votes. They find that education, income, race, marital status, church attendance, age, ideological strength, partisan strength, political interest, residential mobility, and gender help to differentiate voters from nonvoters (Ansolabehere and Hersh 2012). Looking at forty decades worth of elections data, Leighley and Nagler (2013) similarly find that income, age, and gender have continually been relevant factors in determining who votes, while they argue that the effect of race is weaker. Table 2 summarizes the literature on misreporting and turnout. Note that the table is not meant to be exhaustive, but rather suggestive of the research on misreporting and turnout.

[TABLE 2 ABOUT HERE]

Of course, if voters and nonvoters are not different with respect to their substantive preferences, then demographic differences are of little importance. There is, however, evidence to support the notion that voters and nonvoters differ in important ways with respect to their political activity. While voters and nonvoters hold somewhat similar social policy preferences, voters have substantially more conservative economic policy preferences (Verba and Nie 1972; Leighley and Nagler 2013). This is a major concern for democratic theorists, but also a concern for pollsters who are tasked with sorting out the electoral preferences of their respondents. If voters' and nonvoters' policy preferences differ, then surely their electoral ones differ, too.

4. Likely voter models

The literature on misreporting and turnout motivate that individual-level clues, such as demographics and political behavior, can help a pollster determine who will vote. Using this information should help pollsters decide which respondents in their sample will actually vote and, in turn, should help make their election predictions more accurate. A likely voter model that is aware of this literature should focus on the changing political factors that put social pressure on individuals to vote after taking into account some fixed, demographic variables that have a persistent effect on turnout. To account for this, most likely voter models include a combination of questions about an individual’s vote intent, voting history, political behavior, and demographics. How this is done varies widely from pollster to pollster, but there are two general approaches: deterministic and probabilistic models (Pew Research Center 2016).

4.1 Deterministic models

Deterministic likely voter models create a likely voter score for each respondent and then create a decision rule to decide whether to include or exclude a respondent’s response when calculating election predictions; as such, this approach is often referred to as a cutoff approach, as pollsters decide which responses to consider and which ones to discard for their final predictions. These types of models use either a single question or a series of questions to determine how likely it is that a respondent will turn out on Election Day.

Sometimes they are as simple as asking respondents if they plan to vote – the responses of those who report that they definitely plan to vote or will probably vote are kept, while other responses are discarded. Some pollsters construct composite indices based on a series of questions. They assign a respondent a certain number of points based on their responses to questions about vote intent, how much thought they have given to an election, if they have voted before (in general and in their specific precinct), how often they vote, how closely they follow public affairs, and so on (Pew Research Center 2016). Sometimes these indices can focus exclusively on the combination of past voting behavior and vote intent (Freedman and Goldstein 1996; Murray et al. 2009), while others include questions about the voting process, such as where the respondent’s polling place is (Kiley and Dimock 2009). Some adjust the scores for new voters who will inevitably score poorly on indices that include measures of vote history (Cohn 2016).

From here, pollsters use these scores to take a subset of their original sample. A simple implementation of this is the decision tree proposed by Murray et al. (2009). They consider two survey items: vote intention and vote history. Respondents who report that they intend to vote in the upcoming election and who report that they voted in the previous election are classified as likely voters, while all other respondents are classified as unlikely voters (Murray et al. 2009).

Other pollsters take a slightly more sophisticated approach to selecting their subsample of respondents. Another common way is to take a subset proportional to the projected turnout rate for the upcoming election. This is easiest to show through an example. Suppose that a pollster estimates that the upcoming election's turnout will be 50 percent of eligible voters, a figure calculated using previous elections' turnouts and adjusted for characteristics of the specific election (Pew Research Center 2016). If the population that the pollsters sampled from is all eligible voters, then they simply need to take a subset of the top 50 percent of responses by their vote likelihood index score (Pew Research Center 2016). The electoral preferences of this subset are used to generate an election prediction.

Due to practical constraints, pollsters are not usually able to sample off a list of all eligible voters. Instead, many sample off lists of registered voters. If this is the case, pollsters have to consider how many voting age individuals are registered to vote and adjust their predicted turnout rate for this group accordingly. For instance, the Pew Research Center (2016), using polling data from the 2014 U.S. House elections, estimated that 60 percent of registered voters (a group that comprises 70 percent of adults) would vote, which corresponds to a turnout rate of 42 percent overall.

Simplicity is the greatest benefit of deterministic models. Some respondents end up voting and some will not, so it makes sense to model this reality by including the responses of those who are most likely to vote and excluding the responses those who are least likely to vote. Deterministic models are simple to justify and explain to a broad audience because they resemble how an election works.

On the other hand, cutoff approaches suffer from losing information from respondents who are just slightly below the threshold. Many of the respondents whose responses are not considered in the pollster's election prediction do end up voting, so if their electoral preferences are different from the respondents' whose responses are used, bias is introduced into the estimate (Pew Research Center 2016). If it is reasonable to assume that the preferences between these two groups are not too different, then this is not a big issue. But this assumption does not usually hold.

4.2 Probabilistic models

Probabilistic models take advantage of the same series of questions but, instead of determining a cutoff point for which responses to keep, they assign responses a weight: responses from those who are more likely to vote are weighted more heavily than responses from those who are unlikely to vote, but all are included in the election prediction. (Alternatively, these survey weights can be used in a cutoff approach, much in the same way that vote likelihood index scores are used; a pollster can specify that they expect a turnout rate of, say, 60 percent, and partition the top 60 percent of respondents based on these vote likelihood weights to compute their election predictions.)

How are these survey weights calculated? Using a combination of historical data and the data at hand. Pollsters use the same predictors considered in the deterministic models to create a likelihood of voting for each respondent, although it is also possible to include a huge number of demographic variables in the creation of these weights, especially when using machine learning methods that are better equipped to handle predictions from large data sets (Pew Research Center 2016). Then, using data from previous elections, pollsters model a relationship between each of these variables and whether or not an individual voted (Pew Research Center 2016). This model is applied to the current data to create the survey weight, with the assumption that "expressions of interest, past behavior and intent all have the same impact" (Pew Research Center 2016).

Two common methods to generate these weights are logistic regression and random forests (Pew Research Center 2016). Logistic regression, a common statistical modeling tool, transforms the binary vote variable into a continuous variable in order to fit a regression model on the data. The model is trained on historical data, with various predictors as the independent variables and validated vote as the dependent variable, and is used to calculate the weights. Random forests algorithms use a large number of decision trees, each fitted to random subsets of the data and provided with random subsets of all of the available predictor variables at each possible split, to make predictions about who will vote. Predicted probabilities of voting can be extracted from the output of this algorithm.

An additional method, although it is unclear if any pollsters use it in practice, is to borrow vote propensity scores calculated by private voter file firms, such as Catalist or TargetSmart (Pew Research Center 2016). These firms sell large amounts of individual-level data from a variety of sources to campaigns and, more recently, to academic institutions (Ansolabehere and Hersh 2012). In addition to the raw data they provide they maintain a number of proprietary models to predict characteristics about individuals that are not available through the voter file or marketing data sets they obtain. Although the full details of the model are not typically available, the vote propensity scores computed by Catalist use demographic characteristics, commercial data, and vote history.[1] These scores can be used in the same way that the weights from logistic regression models or random forests are used.

The benefits of probabilistic models (that are used for weighting) relative to deterministic ones are that pollsters get to consider more information, as probabilistic models do not sacrifice the information from respondents projected to be marginal voters. By simply weighting unlikely voters' responses less, their preferences are still considered, just to a lesser extent. Using probabilistic models, whether in a cutoff of weighting approach, can be problematic relative to models that do not use historical data at all, as this type of model assumes that the upcoming election will be similar to past ones and that the same set of variables is relevant for predicting turnout from one election to the next. These can be risky assumptions to make. Finally, while more complex models may be better for describing who votes in a specific election, they may make poor out-of-sample predictions and may be hard to justify or explain to the general public.

In the next section I explore a new direction for likely voter modeling: the integration of structural election forecasting models. Structural election forecasting is an area that has the potential to contribute to likely voter modeling but has not yet been used. Before I begin describing the data and methods I use, I also discuss the principles that I consider as I construct and evaluate various likely voter models.

5. Structural election forecasting

In addition to the polls, the other major way that election results are predicted using tools from political science is through structural election forecasting. Structural models, often referred to as models of the fundamentals of an election, seek to predict the two-party vote share of either

[1] Professor Brian Schaffner has access to an academic subscription to Catalist, which he has generously allowed me use for this project.

the Democratic and Republican candidate using data that is available early in the campaign, usually around or slightly after primary elections occur. These forecasts take advantage of economic and public opinion data in conjunction with theories of voter behavior, such as retrospective voting, in which voters out officials who are doing a bad job and keep those who are doing a good job, to suggest that the set up of an election provides information about the outcome. In order for this to happen, at least one of two conditions must hold: (1) voter preferences change based on the set up of an election, or (2) who turns out changes based on the set up of an election. If the latter is true, then it follows that models of the fundamentals would also provide pollsters' with good information for their likely voter models.

One of the most prominent structural election forecasting models is Alan Abramowitz's Time-for-Change model, which uses the popularity of the incumbent President (measured by Gallup Poll's net approval rating at mid-year), the state of the economy (measured by the real annualized gross domestic product growth during the second quarter), and the length of time, in terms, that the President's party has controlled the Oval Office to predict the two-party vote share for the candidate from the incumbent's party (Abramowitz 2008). It was also updated in 2012 to include a measure of political polarization (Abramowitz 2012). To forecast a specific election, the model is typically trained on previous election data and then this specification is applied to the data for the election at hand. In 2008, for instance, Abramowitz trained a model on election data from 1992 to 2004 and then predicted the Republican Party's vote share in the 2008 election.

What is interesting about Abramowitz's model from the perspective of a pollster is not so much about the coefficients of the regression that was fit or about the accuracy of the prediction (although it was relatively accurate in 2008), but about the choice of variables that are considered. Abramowitz considers these variables because they capture the mechanism of retrospective voting: if the economy is bad, the public dislikes the performance of the sitting President, and there is a high level of polarization, then when that President's second term is up the voters will vote them out of office. He also makes the argument that these variables perform better relative to other measures of the economy, Presidential job approval, length of time the current party has been in power, and polarization.

Abramowitz's model is one of many but his variables are carefully selected to provide information about the election outcome and because of this there is reason to believe that they also provide information regarding the composition of the electorate. Research suggests that some externally-focused variables – such as church attendance and the racial composition of one's area – play a role in turnout and misreporting voting behavior (Bernstein et al. 2001; Ansolabehere and Hersh 2012). Currently, it is possible that likely voter models are missing out on this information. Catalist's documentation for its vote propensity models, for example, indicates that they do not intend to capture the "idiosyncrasies of the coming election, such as evolving enthusiasm levels, charisma of candidates, and changing laws." The inclusion of structural election variables into these models could capture some of these idiosyncrasies.

6. Principles of model building

Before I describe the data and methods I use, it is important to introduce a few considerations that I bear in mind as I evaluate different models. The first is parsimony. A good model should be complex enough to capture the underlying process it estimates but not overly complex or bogged down with irrelevant variables. This is especially important for likely voter models, which pollsters should be able to explain to the general public. While evaluating models, I will not only keep in mind how accurate they are but also how simple they are to implement and to explain. A model that cannot be reasonably implemented by a pollster or explained to a reader is not particularly useful.

A second consideration is that these models should be used to give probabilistic rather than deterministic estimates whenever possible. Among others, Nate Silver of FiveThirtyEight has argued that election forecasting models often suffer from “errors of interpretation and communication,” including a failure to express a model’s prediction with the uncertainty that it merits (2017). For likely voter modeling, expressing estimates probabilistically corresponds to evaluating estimates under a range of different turnout scenarios.

Finally, I will evaluate likely voter models not only on their performance predicting the national result but also on their performance predicting the result in each state. One criticism that has been levied against pollsters in the past is that they “typically use a single likely-voter model for the entire country” even though “political science research has shown that state-level factors such as registration requirements, early voting rules, and competitiveness can affect an individual’s likelihood of voting” (Hillygus 2011). In the next section I describe why the data I use for my analysis is particularly well suited for constructing state-level likely voter models.

7. Methodology

7.1 Data

In order to build and assess the performance of likely voter models, I use CCES surveys fielded during presidential (2016, 2012, and 2008) and midterm (2014 and 2010) election years over the past decade[2]. In election years, the CCES survey is fielded in two waves: one collected a few weeks before the election and one collected shortly after the election. The pre-election wave

[2] I use the cumulative CCES file and merge in the items that are not included in the cumulative file by respondents’ unique identifying numbers for each year.

closely resembles a poll taken late in the campaign (the survey is fielded in late September and October leading up to a November election), and thus, the questions asked on it act as good approximations of the information that a pollster would have to make one of their final estimates. While the accuracy of respondents self-reported vote choice and vote likelihood may continue to improve even closer to Election Day, the CCES is still a suitable poll proxy, since polls taken during the last few weeks of the campaign are historically the most predictive of actual election results (Gelman and King 1993).

There is even more reason to believe that the CCES is a desirable data source for the project of modeling likely voters. First, since 2008 the CCES has used vote validation, which is considered the gold standard for studying individual-level election turnout. After the election, CCES respondents are matched into Catalist's national voter file database, which consists of over 240 million unique voting-age individuals across the United States that the organization has compiled by collecting voter registration records from each state and combining these records with commercial records purchased from data aggregators. Their database allows their clients, like the CCES, to identify with a high level of accuracy which individuals have a record of voting in a certain election and which individuals do not (Ansolabehere and Hersh 2012). A validated vote record for each respondent is the important dependent variable for the logistic regression and random forest approaches to likely voter models, which require that a relationship between a number of covariates and turnout be estimated from historical data and then applied forward to new data.

Second, the CCES is a nationally representative and large-N survey, which allows for generalizability of the results of this project. The CCES uses a sample matching process that leverages American Community Survey targets for "age, race, gender, education, marital status, number of children under 18, family income, employment status, citizenship, state, and metropolitan area;" Current Population survey data on voter registration, turnout, and vote choice; and Pew U.S. Religious Landscape Survey information on religion, religiosity, news interest, and partisan and ideological affiliation (Ansolabehere and Schaffner 2017). Additionally, the surveys compensate for any additional lack of coverage by employing a survey weight that considers "age, gender, education, race, voter registration, ideology, baseline party ID, born again status, political interest, plus their interactions" (Ansolabehere and Schaffner 2017). As a result, the focus of this project can rest squarely on the performance of likely voter models instead of the quality of the sampling method and weighted election prediction calculations can be made without any additional work.

Furthermore, the CCES collects a fairly large sample from each state, allowing for robust conclusions about the effectiveness of various state-level modeling approaches in even the smallest of states. The state with the fewest total respondents over the five years I consider (Wyoming) still has nearly 500 valid observations. The overall sample sizes for each year's survey are displayed in Table 3. Ideally, I would have access to data from more election years, but vote validation has only become widely available in the last decade or so (Catalist was formed in 2006) and, since self-reports are notoriously unreliable, taking advantage of data sources that validate turnout are far preferable to those that do not. Since the CCES has so many respondents, it is still possible to conduct a robust investigation of what covariates are correlated with voting, especially when data from the 2010 and 2014 midterm elections is included.

[TABLE 3 ABOUT HERE]

Finally, the CCES's Common Content provides a wide range of demographic and attitudinal questions that may be useful for identifying likely voters based on literature about misreporting and turnout. The survey items this project takes advantage of operationalize vote intent, vote choice, vote history, voter registration status, interest in politics, age, gender, education, race, income, partisanship, religiosity, marital status and residential mobility. Even though there is theoretical justification for their utility, racial composition of district and political activism items are not included in my analysis because they are not widely available in the years I consider. See the Appendix A1 for CCES question wordings. While not every pollster asks all of these questions, their inclusion on the CCES provides me with a richer data set than if I were to look at data from one pollster.

Additionally, I consider structural election variables in a few of my likely voter models. I follow Abramowitz's Time-for-Change model (2012), as revised to not only include presidential approval, economic growth, and whether or not an incumbent ran, but also a measure of the level of polarization. Presidential net approval is taken from the final Gallup poll in June in both presidential and midterm election years. I use the annualized real GDP growth in the second quarter of the election year as reported by the Bureau of Economic Analysis. I slightly tweak Abramowitz's incumbency and polarization variables to fit midterm election years as well as presidential election years. Instead of using whether or not an incumbent ran, I consider whether the incumbent is from the same party as the President. In presidential election years this is coded as 1, but in midterm election years this can be coded as 1 or 0 depending on the party of the incumbent. I also slightly adapt the indicator variable for the level of polarization. In presidential years, it is coded 1 if an incumbent is running or there is an open seat where an incumbent has a net approval rating over 0, and coded 0 if no incumbent is running and if the incumbent has a net approval rating less than 0 (Abramowitz 2012). In midterm years, this variable is coded 1 if the House incumbent is from the same party as the President or if they are from different parties and the President has a net approval rating over 0, and is coded 0 if the House incumbent and President are from different parties and the presidential net approval rating is less than 0.

There are, however, still a few limitations to the CCES data. Most notably, the vote preference question is a noisy estimate of how people actually voted. On most polls, if a respondent reports that they are unsure how they will vote in an upcoming election they are asked a follow-up question regarding which candidate they are leaning toward. Since leaners behave similarly to those with stronger preferences, the preferences of the two groups can be combined to reduce the effects of differential nonresponse and get more accurate estimates (Bruce et al. 1986). The CCES does not ask a follow-up question, so the preferences of respondents who report that they are not sure who they will vote for are completely lost, even if they are leaning toward one of the candidates. This is especially problematic since, while Clinton leads among the overall CCES sample that reported a preference, Trump holds a sizeable 15-point lead among those who responded that they were unsure during the pre-election wave but reported a vote choice during the post-election survey.

There is also a small amount of inconsistency in the operationalization of the vote history variable. For 2012 and 2016, the indicator variable is coded 1 if a respondent reported that they voted in the previous presidential election and 0 if they were unsure, did not recall, or reported that they did not vote. In 2010 and 2014, it is coded similarly, using a respondent's vote history in the previous presidential election rather than the previous midterm election. In 2008, respondents were not asked about their voting behavior in the 2004 presidential election, so whether or not they reported that they voted in a primary election or caucus in 2008 is used as a proxy. Although this is not a perfect substitute, there is substantial correlation between the two questions in the 2016 sample, which was asked about voting in the 2012 presidential election and the 2016 primary election. In 2016, over 72 percent of respondents who voted in the 2012 presidential election also voted in the 2016 primary, while over 78 percent of those who did not vote in 2012 also did not vote in a primary or caucus in 2016.

I also include some details on how I recode any remaining CCES variables I use so that my analysis may be replicated. In 2012 and 2014, respondents were given the option for the vote intent item to respond that they planned to vote before Election Day; these respondents are recoded as definitely intending to vote in that election, in line with how the question is coded in the other years. For the political interest item, respondents who reported that they did not know how often they follow what is going on in government and public affairs were placed in the same category as those who were "[h]ardly at all" interested. While the family income variable on the CCES has many categories, I collapse the categories into four: under $40,000, $40,000 to $100,000, over $100,000, and prefer not to say. Similarly, the six marital status categories are combined into married, single, and other. Using the seven-point partisan identification Likert-scale I compute a four-point index of partisan strength; strong Democrats and Republicans are coded with 1s, not very strong Democrats and Republicans are coded as 2s, Democrat and Republican leaners are coded with 3s, and pure independents and those who are not sure are coded with 4s. Race categories are collapsed to white, Black, Hispanic, Asian, Native American, and other. All respondents with either no record of voting or who had no voter file match are treated as validated nonvoters, while those who are matched successfully are considered validated voters. All coding decisions are made in hopes of leveraging the granularity that the rich CCES dataset has to offer while also making my results accessible to pollsters who likely provide less response options than the CCES. I also make a few adjustments based on lack of variation in the data. In South Dakota, there were no Asian CCES respondents until 2016, so their race is categorized as other so that the training (all years except 2016) and test (2016) sets have identical factors. Additionally, I drop the incumbency variable from models in Massachusetts, Connecticut, New Hampshire, Vermont, Rhode Island, and D.C. because there is no variation in this variable between the training and test sets for these states.

Between the five CCES surveys I consider, there are 263,535 observations. After removing miscoded and missing values, my final data set contains 259,940 observations. Of the 3,595 observations I lose, 225 are dropped from 2008, 688 from 2010, 1,694 from 2012, 497 from 2014, and 491 from 2016. I choose to use this data set throughout to maintain consistency, even though it would be possible to use the full 2016 CCES for the baseline models I describe in the next section. Due to this, the results I present in this paper may vary slightly from what one would find if they used the full 2016 CCES, especially for the baseline models.

7.2 Modeling approach

My likely voter modeling process is broken into two general steps. In the first, I look at all responses as if they were taken as a part of a national poll and use them to create national likely voter models. In the second step, I treat respondents from each state as if their responses were fielded as a part of a state-level poll for that state and construct likely voter models for each state using the respondents that fall under that jurisdiction. In each step, I evaluate how well the models predict individual-level turnout and how well the models can be used to predict accurate estimates of election result. For the purposes of this project, I choose to use data from 2008, 2010, 2012, and 2014 as my training data, where applicable, while all my models are evaluated on 2016 data.

Within each of these steps, I employ both cutoff and probabilistic approaches. First, I begin by using the vote intent question to construct a cutoff likely voter model. I take four different subsets of the 2016 CCES sample: one for respondents who reported that they already voted or reported that they would definitely vote; one for respondents who reported that they already voted or who reported that they would probably or definitely vote; one for respondents who reported that they already voted, would probably or definitely vote, or who were undecided; and one for all respondents. Using each of these subsets, I then use the validated voter status of each subset to compute the true positive rate (rate at which predicted voters are validated as voters) and true negative rates (rate at which predicted nonvoters are validated as nonvoters). Note that I do not use sample weights for this calculation, as I am primarily interested in how well I predict which actual respondents vote. Finally, I compute Clinton's projected margin of victory over Trump for each of these likely voter subsets. The second model I consider simply adds vote history – a subset is created for respondents who reported that they voted in 2012, which is further partitioned by the vote intent categories. Respondents who were not old enough to be eligible to vote in 2012 – i.e., those who are younger than 22 years old in 2016 – are just evaluated on their response to the vote intent question.

The third cutoff model is a reformulation of Pew's Perry-Gallup index. The Perry-Gallup index is composed of seven questions that capture how much thought a respondent has given the upcoming election, whether the respondent has ever voted in their current district, how closely the respondent follows government and public affairs, how often the respondent votes, the respondent's vote intent for the upcoming election (there are two questions on this), and the respondent's vote history (Pew Research Center 2016). Respondents are assigned points based on their responses to those questions (for instance, one point is awarded if a respondent reports that they always or nearly always vote (Pew Research Center 2016). Their scale runs from zero (least likely to vote) to seven (most likely to vote). See Appendix A2 for question wordings and how points are assigned.

The CCES does not contain all of the questions that are used in the Perry-Gallup index and the question wording varies for the items that do appear. Instead, I use vote intent, vote history, and political interest from the CCES to create a version of this index. Together these questions capture five of the seven items on the Perry-Gallup index; the one dimension they do not capture is self-reported historical voting behavior, as the CCES does not ask about whether a respondent has voted in their district before or about their voting frequency. I assign respondents points based on the follow criteria: two points for those who reported that they already voted (early or absentee) in the 2016 general election and those who report they will "definitely" vote, and one point for those who will "probably" vote in the election. Respondents who reported that they voted in the 2012 general election are awarded one point. Those who follow what is going on in government and public affairs "most of the time" are given two additional points, while those who follow "some of the time" are awarded one additional point. I make two further adjustments. Since Pew samples off a list of registered voters, I give respondents who report that they are registered to vote one point. Further, since respondents who are younger than 22 would not have had the chance to vote in the previous election, they are given one additional point. The minimum score in this version of the Perry-Gallup index is zero while the maximum score is six. As I do with the vote intent and vote intent plus vote history models, I create likely voter subsets based on these scores and compute true positive and negative rates for predicted individual-level turnout and calculate election predictions.

While I am interested in how each of these simplistic models performs, I include them largely as baseline models for more complex models. The next class of models I consider uses logistic regression. Using data from 2008, 2010, 2012, and 2014 as training data, I build models that relate three different sets of predictor variables to a respondent's validated vote record. The model can then be applied to new data – in this case the 2016 CCES sample – to get predicted probabilities of voting for each respondent. The first set of predictors I consider is merely the variables that are included in my version of the Perry-Gallup index: vote intent, vote history, political interest, as well as voter registration status and age. So that the eligibility to vote in the previous election variable is defined similarly to how it is defined in the index, I create an indicator variable that is coded 1 if a respondent was old enough to vote in the previous election and 0 if the respondent was not old enough to vote in that election. To further capture the relationship between the vote intent and vote history items, I include their interaction term.

The second set of variables consists of the Perry-Gallup index variables and over a half dozen demographic variables that are both theoretically tied to misreporting voting behavior or turnout and available on the CCES. Those variables are age, race, education, family income, partisan strength, religiosity, marital status, and residential mobility. Note that this set of variables does not use the eligibility to vote indicator variable since it includes age. The third set of variables adds structural election variables from Abramowitz's Time-for-Change model (presidential approval, economic growth, the incumbent's party, and the level of polarization) to the second set of variables. I include all interaction terms between presidential approval, economic growth, incumbency, and polarization in this logistic regression model.

The final modeling approach is random forests, a powerful machine learning tool that uses a large number of decision trees, each fed with a random subset of the data and a random subset of all possible variables at each split, that can be used to compute vote propensity scores much in the same way that logistic regression can be used. The benefit of random forest algorithms is that, since they randomly sample the data and the available predictor variables, they avoid much of the bias that traditional decision tree approaches encounter, which make them useful for prediction. I use the same three sets of predictor variables with the dependent variable being the binary validated vote. The random forest algorithm outputs predicted class probabilities for each respondent, i.e., each respondent is assessed a probability that they did note vote and another probability that they did vote.

For the national models, I pool all of the observations together as if they were fielded as a part of a national poll. The models' performance is evaluated using all of the 2016 CCES sample and the logistic regression and random forest models are trained using all of the data from 2008, 2010, 2012, and 2014. For the state models, though, I evaluate each type of likely voter model 51 times (for all 50 states and the District of Columbia) using only data from that state. Further, for the logistic regression and random forest approaches, I train each state's model only with historical data from that state to isolate unique characteristics of each state. For instance, the likely voter model that is applied to 2016 CCES respondents from Texas is only trained on data from other respondents from Texas.

8. Results

8.1 Vote intent

I begin by looking at three simple baseline likely voter models: vote intent, vote intent and vote history, and a variation of the Perry-Gallup index. The ideal scenario for a pollster would be that one of these three methods produces election estimates that are accurate, as they are easy to implement in any election and even easier to justify to a wide audience. But, as expected, using vote intent alone is not particularly effective in differentiating voters from non-voters, at least in a way that allows pollsters to predict the eventual result with minimal error. This holds true when this approach is used nationally and state by state.

In 2016, 49,527 respondents reported that they would definitely vote, or greater than 77 percent of the CCES sample from that year. Roughly 7.5 percent of respondents checked that they would probably vote while 2.4 percent indicated that they had already voted via early or absentee voting. Just over eight percent said that they would not vote in the 2016 general election and 4.8 percent reported that they were undecided. While these numbers seem high – nearly 90 percent of respondents in the 2016 CCES appear to vote -- it is important to keep in mind that respondents misrepresent their voting behavior, as only 53 percent of the 2016 CCES had a validated voting record once the sample was matched into Catalist's database.

It is also important to keep in mind that 2016 was not an anomaly in this respect. In the other two presidential election years in which the CCES used vote validation, respondent misreporting followed similar trends. Over 82 percent of respondents in 2012 and just over 80 percent of respondents in 2008 said they would definitely vote, but only 67 percent of 2012 respondents and 68 percent of 2008 respondents were validated voters.

Narrowing down the definition of a likely voter from all respondents to those who have already voted or who say they will definitely vote does improve the rate at which actual, validated voters are predicted to be voters using this model, which is also known as the true positive rate. When the sample is restricted to these respondents, 62.5 percent of predicted voters are validated voters (conversely, 84.3 percent of predicted nonvoters are not validated as voters, which is known as the false positive rate). As the definition of who a likely voter expands to also encapsulate respondents who say they will probably vote, the true positive rate falls to 59.6

percent, while the true negative rate climbs to 91.4 percent. When undecided voters are included in that definition the true positive rate decreases to 57.2 and the true negative rate rises to 95.2 percent. In general, this trend makes sense, as broader definitions of likely voters should predict that more nonvoters vote while less actual voters should be predicted to be nonvoters.

Of course, it will never be possible to perfectly predict which respondents will vote. But if the preferences of predicted voters closely resemble those of all voters, the pollster's job is complete. Using each definition of likely voters I calculate the share of votes that each candidate would have received if only that subset turned out to vote.[3] Since the race was primarily a contest between Trump and Clinton, I focus on the difference between these two candidates' vote shares. Figure 1 shows Clinton's predicted margin of victory over Trump, using the overall weighted vote shares of each candidate, as well as the other and undecided options.

[FIGURE 1 ABOUT HERE]

Figure 1 shows that, for any definition of who a likely voter is that relies solely upon the vote intent item, the estimates generated from that likely voter subset are fairly inaccurate. Defining likely voters just as those who said they have already voted or those who report they will definitely vote, Clinton's predicted margin of victory over Trump is just over 6.5 points. This is quite a way off from the 2.1 points that she actually won by nationally. But, it is also over 3.5 points off her margin of victory among all validated voters in the 2016 sample, which I use as a proxy for where the race actually stood when the survey was fielded. That mark, +2.9 points in favor of Clinton, is shown by the red dashed line in Figure 1.

Next, instead of treating all observations as if they came from a national poll, I group together observations from the same state and treat all the respondents from each state as if they

[3] The candidates listed on the 2016 CCES were Donald Trump, Hillary Clinton, Gary Johnson, Jill Stein, as well as an option for all other candidates and undecided respondents.

came from a state poll. In 2016, every state except Wyoming (N = 99) had over 100 respondents on the CCES. In the larger CCES sample I use to train the logistic regression and random forests models, every state has over 500 observations across the five election years I consider (except for Wyoming, which has 486). These large sample sizes provide me the ability to make robust conclusions about the effectiveness of various types of likely voter models.

At the individual level, a more stringent definition of who a likely voter is results in higher average true positive rates across all states, as well as lower average true negative rates. As the definition gets more expansive, including not only those who reported that they already voted or would definitely vote but also those who said they would probably vote or were undecided, the average true positive rate decreases while the average true negative rate increases. As before, whether a pollster accurately forecasts whether any set of individuals actually votes pales in comparison to the importance of identifying a subset of voters whose preferences reflect those of the population at large.

Figure 2 shows Clinton's predicted margin of victory over Trump (x-axis) plotted against her lead among all voters (y-axis) in all 50 states and the District of Columbia. Points below the dashed diagonal reference line correspond to states in which Clinton's projected lead is overestimated with respect to the preferences of validated voters in that state at the time the survey was fielded while points above the line indicate that her lead was underestimated. For each version of the vote intent model that I consider, the majority of points fall below the reference line. Using only vote intent to define the likely electorate, support for the Democratic candidate would have been overestimated in a majority of states, regardless of which specific definition was used. The average error was highest when likely voters were defined as those who already voted and those who would probably or definitely vote (2.6 points, mean squared error =

46.5) and lowest when all respondents were used to calculate estimates (2.3 points, mean squared error = 49.2).

[FIGURE 2 ABOUT HERE]

8.2 Vote intent and vote history

Next I add vote history into the likely voter model. The analysis follows the same pattern as the previous section, except I further subset the data to only take voters who reported that they voted in the 2012 general election and not those who did not recall or who reported that they did not vote. Those who were not old enough to vote in the previous election are evaluated solely on their response to the vote intent question. Roughly 75 percent of respondents in 2016 reported that they voted in the 2012 general election. Of the respondents who reported that they voted in 2012, over 90 percent responded that they would definitely vote in the 2016 election. Misreporting is apparent again, as only 62.4 percent of 2016 respondents who reported that they voted in 2012 ended up voting in the 2016 election. Furthermore, among the 2016 respondents who were too young to vote in 2012, a small fraction (less than 5 percent) erroneously reported that they did vote in 2012.

Adding this criterion to the model, however, does not improve upon the performance of the model that only uses vote intent. The true positive rate ranges from 62.9 percent when likely voters are defined as those who reported that they voted in 2012, who already voted or who said they would definitely vote in 2016 to 60.5 percent when all respondents who said they voted in 2012 are considered. Meanwhile, the true negative rate slightly increases from 78.5 to 77.8 percent as I move from the most restrictive to the least restrictive definition. Figure 3 shows Clinton's predicted margin of victory over Trump by different vote intent responses for both those who reported that they voted in 2012 and all respondents. Using the vote history criterion pushes the estimated margin between 0.4 and 1.1 percentage points further away from the actual margin among validated voters in the sample. This suggests that, once vote intent is used in this simplistic fashion, adding self-reported vote history does not add any predictive power.

[FIGURE 3 ABOUT HERE]

At the state level, it also seems reasonable to try to remedy the persistent Democratic bias across states by incorporating vote history into the model. I further subset each of the vote intent subsets by whether or not respondents reported that they voted in the 2012 general election. This small tweak does affect how well each model predicts whether or not a respondent voted in the 2016 general election. After this adjustment, there is little difference in the distribution of true positive or true negative rates across different vote intent categories across states. In other words, once vote history is incorporated into the model, defining the likely electorate more narrowly or more broadly does not change how well the model predicts which respondents actually vote.

When this vote intent and vote history model is used to generate election predictions, it performs worse than the models that only use vote intent. Roughly three-quarters of states produced election predictions that overstated Clinton's chances with respect to her lead among validated voters in the state at the time, which can be seen in Figure 4. Figure 4 plots Clinton's predicted margin of victory over Trump among self-reported 2012 voters and for different vote intent categories against her lead among all voters in each state. Points below the dashed diagonal reference line correspond to states in which Clinton's projected lead is overestimated with respect to the preferences of validated voters in that state at the time the survey was fielded while points above the line indicate that her lead was underestimated. For each vote intent category, the vast majority of points fall below the reference line. The average prediction error for each of these models is also greater than it was when vote history was not factored into the mix. When likely voters are defined as those who already voted or those who would definitely vote, the average error increases by 0.4 points when vote history is included; when those who will probably vote are included, adding vote history increases the average error by 0.25 points; and when undecided voters are included in the likely electorate, the mean prediction error increases by nearly 0.7 points once the likely electorate is partitioned by vote history. In the whole sample, using voting in 2012 as the only criteria to define likely voters increases the average state-level prediction error by almost 0.9 points.

[FIGURE 4 ABOUT HERE]

8.3 Perry-Gallup index

For the third baseline model, I consider a variation of the Perry-Gallup index. Respondents receive points toward a likely voter score based on their responses to questions about vote intent, vote history, political interest, and voter registration status, as well as for being old enough to vote in the previous presidential election. Those who are the most likely to vote receive a score of 6 while the least likely to vote receive a score of 0. The sample is then partitioned into a likely voter subset using these scores and individual-level turnout and election predictions are evaluated.

Only 62 of the 64,109 respondents (less than a tenth of a percent) received the highest score. The majority of respondents fell into the 3-5 buckets: 41.7 percent received a score of 5, 23.1 percent received a score of 4, and 16.5 percent received a score of 3. Roughly 17 percent of respondents received a score of either 2 or 3, and less than 1.5 percent received a 0, the lowest score.

The top category does not do a particularly great job of predicting turnout on an individual-level: the true positive rate is slightly greater than 30 percent, which is worse than the true positive for the entire sample (53 percent), and the true negative rate is under 50 percent. Once respondents that received a score of 5 are added, the true positive rate drastically improves to 68.2 percent and the true negative rate reaches 58 percent. Adding respondents who received scores of 4 brings the true positive rate down to 64.2 percent but brings the true negative rate to roughly 68 percent. Similarly, the true positive rate goes down to 60.8 percent once 3s are added while the true negative rate rises to 81.5 percent. Finally, adding respondents who received scores of 2 to the likely voter definition brings the true positive rate down to 57.6 percent and the true negative rate to nearly 90 percent. This follows the same trend as the vote intent and the vote intent and vote history models: more restrictive definitions of likely voters do a better job of predicting actual voters and a poorer job of predicting nonvoters while more lenient definitions do a worse job predicting validated voters and a better job of predicting validated nonvoters.

This method does a slightly better job at generating accurate estimates than the previous two methods, using a reasonable cutoff point on the scale. While Clinton's margin of victory is +18.2 points when only respondents who received a score of 6 are considered likely voters, when those who received 6s or 5s are considered likely voters her predicted margin of victory over Trump is 1.7 points (1.2 points off the mark for validated voters in the sample), and when those who received 6s, 5s, or 4s are considered likely voters the margin is just shade over 5 points (2.1 points off), as shown in Figure 3. Once any more categories are added to the likely voter subset, though, her projected lead grows to over 6 points.

Moreover, the 5s and 6s or 4s, 5s, and 6s categories make practical sense to use because they suggest turnout rates that are in the ballpark of what turnout rates should be for presidential elections and, indeed, what the turnout rate actually was in 2016. Defining likely voters as those who receive 5s and 6s provides a subset of roughly 42 percent of the sample, while expanding the definition to also include 4s puts the subset at 65 percent of the sample. In 2016, the actual voting age population (VAP) turnout rate was 54.7 percent (McDonald 2018).

[FIGURE 5 ABOUT HERE]

At the state level, I follow the same pattern. Here, though, I do not evaluate respondents who score 6s on their own (the most restrictive likely voter category considers both 5s and 6s) because many states have so few respondents who receive a 6 on the index. The true positive and true negative turnout trends follow the same pattern that they follow in the above models. When respondents who score 6s and 5s on the index are considered likely voters, the distribution of states' true positive rates is centered at roughly 74 percent. As respondents who receive lower scores are added to the likely electorate, the true positive rate gradually decreases; it falls to 70.2 percent when 4s are added, to 66.5 when 3s are added, and to 63.3 when 2s are added. Conversely, the percent of respondents who are not in the likely voter subset and who do not end up voting increases as more index scores are added to the likely electorate. When just 6s and 5s are considered likely voters, only 56.1 percent of predicted nonvoters are actual nonvoters while, when all respondents who receive above a 0 are considered likely voters, nearly 90 percent of predicted nonvoters do not actually vote.

As was the case for national models, this index-based likely voter model also does a better job of estimating major candidate vote shares than the vote intent or vote intent and vote history models. When 6s and 5s are considered likely voters, the model actually overestimates Trump's support in 32 states. The average error is +2.3 point in favor of Trump under this model. But, when 4s are added to the likely electorate, the model only slightly overestimates Clinton's support, as the average difference between the predicted margin in a state and the margin among validated voters in that state is 1.4. As more and more index scores are added to the likely voter definition, Clinton's lead continues to be overestimated, but it plateaus around an average error of 2.2 to 2.4 points in her favor. Figure 6 shows Clinton's projected lead over Trump (for different likely voter index score groups) against her lead among all voters for each state. Points below the dashed diagonal reference line correspond to states in which Clinton's projected lead is overestimated; points above the line indicate that her lead was underestimated in that state. As the likely voter definition gets more expansive, more and more points shift below the dashed reference line, suggesting that more lenient likely voter definitions lead to overestimates of Clinton's lead.

[FIGURE 6 ABOUT HERE]

8.4 Logistic regression

Next I turn to the two more complex modeling approaches: logistic regression and random forests. In each case, I consider three sets of predictors: variables from my version of the Perry-Gallup index; all of the variables from the first model plus nearly a dozen demographic variables that are theoretically tied to turnout and misreporting voting behavior; and then all of the variables from the second model plus structural election variables. A model is trained on data from 2008, 2010, 2012, and 2014, relating each set of variables to whether or not an individual actually voted, and then used to compute a vote propensity score for each 2016 respondent.

I first look at how well each model predicts turnout at an individual level. To do this, I consider each possible turnout rate X, for X $\in$ [1, 100], and then divide the data into a likely voter subset by taking the top X percent of respondents based on their vote propensity scores. If multiple respondents at the cutoff point have the same propensity score, they are all are counted as likely voters. From each subset, true positive and true negative rates are calculated (see Figure 7). For each set of predictors the two rates follow a similar trend: low turnout rates are tied with low true positive rates and, relatively, high true negative rates, while higher turnout rates are associated with a steady increase in the true positive rate and a steady decrease in the true negative rate.

[FIGURE 7 ABOUT HERE]

While it is unlikely that a pollster would ever consider a turnout rate of less than 30 or 40 percent for a presidential election, it is shocking how poorly the logistic regression models perform for lower levels of turnout. For lower levels of turnout, respondents who are seen as extremely likely to vote with respect to historical trends are identified as voters, but in 2016 these are not the types of respondents that actually vote as frequently. In fact, there is no likely voter subset produced by any of the three models that features a true positive rate above the validated turnout rate for the entire 2016 sample. Put differently, none of the three logistic regression-based likely voter models do a better job of predicting individual-level turnout than just predicting that all respondents will vote. Additionally, this approach gives respondents in the CCES sample low vote propensity scores. The average score is 0.44 when only Perry-Gallup variables are used, 0.45 when Perry-Gallup and demographic variables are used, and 0.09 when Perry-Gallup, demographic, and structural variables are included in the model.

While this seems problematic, the real job of pollsters is to generate accurate vote share estimates. If the preferences of respondents in the likely voter subset reflect those of actual voters – or at least those of validated voters in the sample – then it is still possible for a pollster to provide accurate election predictions. Using the logistic regression approach, however, pollsters do not do a particularly good job of predicting the overall vote shares for Clinton and Trump. Clinton's predicted margin of victory over Trump, using each of the three logistic regression based likely voter model approaches and evaluated at each turnout rate, is shown in Figure 8. The most simplistic model, which just uses the same variables that are used to compute my variation of the Perry-Gallup index, is the most accurate of the three, but it still misses by quite a bit. Between 40 and 60 percent turnout (a reasonable range for a presidential election), the model predicts a margin of victory for Clinton between 12.2 and 6.7 points, which are 9.2 and 3.8 points off the mark among validated voters in the 2016 sample, the best estimate available for the true standing of the race at the time the survey was fielded. Further, calculating her projected margin of victory using the vote propensity score as a survey weight, in addition to the sampling weight, indicates that she led by 7.7 points, 4.8 points greater than the margin among validated voters.

The two more complex models perform quite a bit worse. The Perry-Gallup index plus all demographic variables model predicts a margin of +17 points for Clinton when the turnout rate is 40 percent and +16.2 points for Clinton when the turnout rate is at 60 percent. These are far from where the race stood at the time the survey was fielded: +2.9 point for Clinton. Under this model the vote-propensity-weighted margin is +10.5 in favor of Clinton, which is nearly three points worse than it was when no demographic variables were considered. Adding structural election variables does not improve or worsen the performance of the likely voter model. Clinton’s projected margin ranges between 15.8 (40 percent turnout rate) and 15.6 (60 percent turnout rate) within the reasonable turnout range, while the margin computed with the sampling and vote propensity weights climbs to 11.5 points. Notably for all three models, there is not a single election prediction made that underestimates Clinton’s support relative to her support among validated voters in the 2016 sample.

[FIGURE 8 ABOUT HERE]

Similarly, for the state-level approach the logistic regression method tends to predict that 2016 CCES respondents in each state have relatively low vote propensity scores. The average vote propensity score for all respondents was 0.44 when a model considering only the Perry-Gallup index variables was used to predict the likelihood of voting for respondents by state. Using this approach, only six states had average vote propensity scores over 0.5. When demographic variables are added to the model the average vote propensity score for all respondents bumps up to 0.45 and a total of eight states have average vote propensity scores over 0.5. Using the most fully specified model produces an average vote propensity score of 0.15 for all respondents, but the averages by state vary widely. I withhold discussion of how this approach predicts individual-level turnout at the state-level because it largely follows the same trend as at the national-level.

In terms of predicting the result in each state's contest, this method produces a consistent Democratic bias for any level of turnout. Figure 9 shows a scatterplot of the error between Clinton's lead among all validated voters in each state and her lead among respondents in each state who are in the likely voter subset. The black dashed line indicates the zero error line and the blue line shows the mean error for all states for each turnout rate. I only look at turnout rates from 30 to 100 percent because, for lower turnout rates, there are so few respondents in some states' likely voter subsets that the estimates are uninformatively noisy; plus, it is unlikely that a pollster would consider a turnout rate below 30 percent. For every level of turnout, the blue line is above the black dashed line, indicating that the average predicted margin across all states was more pro-Clinton than the actual result. Notice that as more variables are added to the model the blue line moves even further away from the black dashed line. When Perry-Gallup index and demographic variables are included, for instance, the average error between the margin among validated voters and among respondents in the likely voter subset at that level of turnout ranges between 8.7 and 9.9 points between 40 and 60 percent turnout (Figure 9 (b)); when structural variables are added into the model, the average error between 40 and 60 percent turnout estimates ranges from 8.2 to 8.7 (Figure 9 (c)).

[FIGURE 9 ABOUT HERE]

When the vote propensity scores are used as an additional weight to calculate Clinton's predicted margin of victory rather than a tool to subset the sample into likely voters, the effect is similar: a Democratic bias across states emerges. Figure 10 shows Clinton's predicted margin of victory weighted by the CCES sampling weight and the logistic regression-based vote propensity weight plotted against her lead among validated voters in each state. Points that lie below the black dashed reference line indicate that Clinton's lead is overstated in that state while points above the line indicate that her lead was understated in that state. In Figure 10 (a), the majority of points fall below the reference line, indicating that even weighting the entire sample's preferences by their likelihood to vote still overstates Clinton's support in a large number of states. This holds true when demographic variables (Figure 10 (b)) and structural variables are added (Figure 10 (c)).

[FIGURE 10 ABOUT HERE]

8.5 Random forests

Finally, I turn my attention to the random forests-based likely voter models. I create three separate models, using the same three sets of predictors that I consider for the logistic regression models, and evaluate them in the same fashion. Figure 11 shows the national true positive and true negative rates by turnout for each of the three versions of the random forests likely voter model. For the random forests models, the two rates follow very different trends. When the turnout rate is lower, and thus the likely voter subset is smaller, the true positive rate is greater than the true negative rate. As the turnout rate increases, the true positive rate gradually drops toward the true positive rate for the entire sample (53 percent) and the true negative rate rises to near 100 percent. Intuitively this makes sense: more restrictive definitions of who likely voters are, aided by historical data to decipher which voters in the current sample are most likely to vote, should identify a lot of actual voters as predicted voters, but may also be prone to identifying some actual voters, who appear to be nonvoter types, as nonvoters. More lenient definitions of likely voters, on the other hand, include many nonvoters in the likely voter subset but do not predict that very many if any, actual voters will not vote. This trend holds true for all three random forests models. As opposed to the logistic regression models, the random forests models do a better job of predicting individual-level turnout than just predicting that all respondents will be voters. In light of this, the decision tree approach appears to work better than the regression approach.

[FIGURE 11 ABOUT HERE]

The indication that the random forests approach is more effective is supported by the election predictions produced by these models, especially the ones that incorporate a wider array of variables. Clinton's predicted margin of victory over Trump, using each of the three random forests-based likely voter model approaches and evaluated at each possible turnout rate, is shown in Figure 12. Within a standard range of turnout rates (40 to 60 percent), using the cutoff approach with vote propensity scores generated by the Perry-Gallup index and all demographic variables model either gives Trump a lead of 5.1 points or puts the race at essentially tied (Figure 10 (b)). For a slightly higher turnout rate, roughly 70 percent, it would be possible to almost perfectly predict where the race stood when the survey was taken. While the cutoff approach still misses the mark by a bit, using vote propensity scores from this model to weight the preferences of all respondents in the sample puts Clinton's lead at 2.5 points – less than a half a point away from what her lead at the time likely was.

[FIGURE 12 ABOUT HERE]

Adding in structural election variables brings the range of reasonable margins a little closer to what the truth was at the time; Trump has a 4.4 point lead with 40 percent turnout and Clinton has a 1.1 point lead with 60 percent turnout. (Nailing the actual mark among validated voters in the 2016 CCES would only require a pollster to use a turnout rate around 66 percent here, which is not unreasonable if a pollster believes that their sample would have a higher turnout rate than the rate among all voting age individuals because survey respondents are generally more politically interested than nonrespondents). Using the vote propensity scores from this model as an additional weight pushes the weighted margin among all respondents slightly further away from the +2.9 mark to +4.1 in favor of Clinton.

The model that is just trained on variables from the Perry-Gallup index indicates that Clinton's lead was at either +1.2 (40 percent turnout) or + 5.2 points (60 percent turnout) in late September and October. Both of these are not too far off Clinton's likely lead of 2.9 points, as computed from the preferences of validated voters, but the lower turnout scenario underestimates support for her while the higher turnout scenario overestimates it. When the vote propensity scores from this model are used as an additional weight to calculate her margin of victory, she leads Trump by 6.2 points.

At the state-level, the random forests approach also outperforms logistic regression across the board. (Again, I withhold discussion of individual-level turnout because it follows the same trend as it does nationally.) While the logistic regression approach suggests that calculating Clinton's lead using any likely voter subset based on any level of turnout produces a substantial

Democratic bias, there is little to no bias using the random forests approach. For any level of turnout – in particular, for any reasonable turnout rate between 40 and 60 percent of the voting age population – Clinton's predicted support is very close to her support among validated voters in the sample in each state. This can be seen in Figure 13, which plots the error between the predicted margin of victory for Clinton and the margin among validated voters for each state by turnout. The dashed black line is the zero error line and the blue line is the average error for all states at each level of turnout. In Figure 13 (a), which looks at the least fully specified random forests model, the blue line is slightly above the dashed black line for reasonable turnout rates, indicating that this approach slightly overestimates Clinton's lead at the time, but the average error is no greater than two points between 40 and 60 percent turnout. Using random forests-based vote propensity scores, created using variables from the Perry-Gallup index, to partition the sample into likely voters and then taking those voters' preferences produces lower error state-level estimates, on average, than using logistic regression-based vote propensity scores.

This holds true for the other versions of the random forests model, which are fed a larger number of predictors. For both of the more fully specified models, state-level estimates are fairly accurate within a reasonable range of turnout; if anything, these models tend to underestimate Clinton's lead at the time across states. The model that considers Perry-Gallup and demographic variables underestimates her support by between 5.6 and 1.9 points, under the most reasonable turnout scenarios. The model that adds structural variables underestimates her lead by between 4.8 and 1.2 points over the same turnout range. This can be seen by the blue line in Figure 13 (b) and (c), which is slightly below the dashed black line that corresponds to no error.

[FIGURE 13 ABOUT HERE]

This approach yields even more accurate results when the vote propensity scores are used to weight the preferences of the entire sample. Figure 14 shows the predicted margin of victory for Clinton, weighted by the sampling and vote propensity weights, plotted against her lead among validated voters in each state. Points that fall below the black dashed line correspond to states in which her support was overstated and points that fall above the line correspond to states in which her support was understated. While the majority of points fall below the line, the points tend to cluster tightly around the line, suggesting that using random forests-based vote propensity scores as an additional survey weight results in, on average, accurate state-level election predictions. In Figure 10, which plots this same result when a logistic regression approach is used, the points spread out

[FIGURE 14 ABOUT HERE]

It should also be noted that the random forests approach generally predicted that 2016 CCES respondents would be more likely to turnout in the 2016 election. Nationally, the average vote propensity score using the model with only the Perry-Gallup index variables is 0.76; the average vote propensity score using the model that adds demographic variables is 0.62; and the average vote propensity score for the model that adds structural variables onto both the Perry-Gallup index items and the demographic variables is 0.66. Using the state-by-state approach, the average vote propensity scores for all respondents are: 0.72 (Perry-Gallup variables), 0.57 (Perry-Gallup and demographic variables), and 0.61 (Perry-Gallup, demographic, and structural variables).

8.6 Summary

My analysis looks at how well various likely voter models perform when applied to data from the 2016 presidential election. I consider five classes of models (vote intent, vote intent and vote history, the Perry-Gallup index, logistic regression, and random forests) at both the national and state levels. For the logistic regression and random forests approaches, I train the models on three different sets of predictors (Perry-Gallup variables; Perry-Gallup and demographic variables; Perry-Gallup, demographic, and structural variables). Each approach is evaluated using the cutoff method and the vote propensity scores generated through the logistic regression and random forests approaches are also tested as weights.

Figures 15 and 16 display the results from all of these efforts. Figure 15 compares the performance of the various national-level likely voter models I consider. For the three baseline models I display the model specification with the lowest absolute error. For the logistic regression and random forests approaches, I consider the range of margins computed using the cutoff approach at reasonable turnout levels (40 to 60 percent) for each of the three sets of predictor variables. Additionally, I include how the model would have performed if the vote propensity scores it computed were used as a weight rather than a sample-partitioning threshold. I compute the error of each approach relative to the preferences of validated voters in the 2016 CCES sample. Points to the right of the dashed vertical line indicate that the approach overestimated Clinton's support at the time the survey was fielded (and are colored blue), while points to the left of the line indicate that Trump's support at the time was overestimated (and are colored red). Figure 16 is similar but looks at the performance of the various models at the state level and measures the average error across all states. For the baseline models in Figure 16, I choose the version that produces that lowest average error across states.

There are only a few models that improve upon simply using the preferences of all respondents to calculate an estimate. Simply considering all respondents as likely voters overestimates Clinton's support by 3.75 points at the national level and by an average of 2.34

points at the state level. The only models that produce an absolute error less than or equal to both the national and state baselines are the Perry-Gallup index model, the random forests model using Perry-Gallup variables (both the cutoff and weighting approaches), and the weighting variation of the both random forests models that consider a larger array of inputs. Using the cutoff approach in conjunction with the two more fully specified state-level random forests models produces estimates that are under the baseline error for the higher end of the likely turnout spectrum but that are over the baseline error when the turnout rate is closer to 40 percent. Overall, the Perry-Gallup index and random forests paired with a cutoff approach are viable likely voter models but training a random forests model on historical data and applying its output as survey weights produces the most consistently low-error estimates.

On the other hand, my analysis indicates that any likely voter model approach that took advantage of logistic regression missed the target nationally and state-by-state, often by shocking margins. For instance, using the cutoff approach along with logistic regression at a national level led to estimates that were anywhere from 3.75 to over 14 points off the true mark. At the state level, these approaches ranged from an average miss of 4.3 points to 9.9 points across all states. Using logistic regression-based vote propensity scores to weight the entire sample's preference did not fare much better.

The model that used just vote intent had mixed results. At the national level, it brought error slightly under what it would have been with no likely voter model. But at the state level, the average error across states using this model was roughly 2.39 points, which is slightly higher than the average error using no likely voter model. Similarly, when self-reported vote history is included, the model is just slightly more error-prone than the method that does not use a likely voter model.

One of the goals of this project was to study how adding structural election information to likely voter models would affect estimates. Once demographic variables are added to the logistic regression- and random forests-based likely voter models, including indicators of economic strength, presidential approval, incumbency, and polarization provide little to no additional predictive information. In Figures 15 and 16, comparing the logistic regression and random forests models that use Perry-Gallup and demographic variables (both for the cutoff and weighting approach) to their counterparts that also include structural variables does not reveal that the additional information moves the estimates any closer to the zero error mark. In some cases, like the national-level logistic regression approach, the more fully specified model performs more poorly when it is used to create vote propensity weights.

[FIGURE 15 ABOUT HERE]

[FIGURE 16 ABOUT HERE]

9. Discussion

In the wake of any election, and especially in wake of the 2016 presidential election, the election polling industry receives marks from political commentators. Often, pollsters' definitions of likely voters are called into question but, due to the lack of publically available information about how individual pollsters model vote likelihood, this criticism lacks empirical support. Other studies have compared how various ways to define likely voters affect trial heat estimates and find mixed results about the effectiveness of the methods they try. This paper aims to extend this line of research, using simple techniques as well as popular and powerful statistical tools to harness the large amount of useful publically available data that exists, and to better integrate probabilistic thinking into the communication of poll results.

It is first important to revisit why polls are even important. Polls not only serve as an important resource for the public and the media to discuss elections, but also as a valuable data source to study shifts in public opinion during an election cycle. And getting it right matters if the industry is to be respected and to maintain public value. In his 1988 book *Pre-Election Polling: Sources of Error and Accuracy*, Irving Crespi writes:

> Even though a pre-election polls is in itself unquestionably a measurement and not a prediction, concluding that even if a poll were conducted immediately before an election, one cannot hope to measure voter preferences accurately enough to approximate election results closely is to impugn the meaningfulness of all polls. If polls cannot achieve such accurate predictability, why should we accept any poll results as having meaning relevant to real life?

Getting it right, however, does not mean making perfect estimates for every election. There is a lot of uncertainty inherent in the likely voter modeling process and it is crucial that this uncertainty is quantified by probability and communicated honestly to the general public. In her essay in "Data and Democracy," Natalie Jackman writes about the difficulty of this type of reporting for election forecasting, but her concern maps well onto likely voter modeling:

> In media forecasting, it's not enough to have a good model. You have to be able to explain it to the audience. This task can be even more difficult than building the model itself. Explaining the uncertainty of probability-based forecasting to the general public is a task that has flummoxed scientists, and particularly weather scientists, for many years. It seems no matter how many times you remind the public that forecasts are based on uncertain probabilities, some people want to read the numbers as completely certain, and then castigate the analysts if the outcome is different from their expectations.

Despite these considerations, a brief examination of the few likely voter models that are publically available suggests that many are unsophisticated, unjustified, ineffective, and lack effective probabilistic communication. A North Carolina poll conducted by Elon Poll a week before the 2016 election, for example, includes only two survey items in their model – vote likelihood and interest in the election (Elon Poll 2016). In light of Bernstein et al.'s conclusion that individually-focused variables (i.e., demographics), as well as externally-focused ones (for instance, the concentration of a minority racial group in the respondent’s district) mediate turnout, Elon Poll’s likely voter model is theoretically inadequate. Elon also ended up calling North Carolina for Clinton, even though Trump went on to win, and the actual margin of victory fell outside their reported margin of error (Elon Poll 2016). This is not to say that likely voter models need to be unnecessarily complex. But when the model fails theoretically and empirically, a new approach is warranted.

On the other hand, there are pollsters using the variation inherent in different versions of likely voter models to report the probabilistic nature of their estimates. The best example of this comes from a New York Times article by Nate Cohn in September of 2016. Cohn provided the same raw data, which came from a Florida poll, to four different pollsters and asked them to provide their estimates for the presidential race in the state. Each pollster used a different method to identify likely voters and to weight the sample. The result was four different estimates for Clinton’s lead in the state at the time. Different pollsters use different methods of modeling vote likelihood and of weighting (these are referred to as a part of a polling institution’s “house effects”) and these decisions impact results.

While the New York Times piece was an exceptional example of data-driven political journalism that roots out the effect of important methodological decisions, such as likely voter models, what is more important is that ordinary pollsters and news organizations reporting on the results of polls incorporate different turnout scenarios based on likely voter models. Ahead of the special election for U.S. Senate in Alabama in November of 2017, SurveyMonkey fielded a poll and, rather than report whether Democrat Doug Jones or Republican Roy Moore was in the lead at the time, SurveyMonkey expressed that the race looked different when you defined likely voters in different ways (Blumenthal 2017). Blumenthal went on to break down estimates generated from the data when likely voters were defined as all registered voters, those certain to vote, those certain and probable to vote, and those who voted in 2014 (or who were age 18-20 and were certain to vote), as well as weighting the preferences of the entire sample by self-reported voting. These different likely voter models, paired with different weighting methods, revealed that the SurveyMonkey data could support the conclusion that Jones would win by nine points, that Moore would win by 10 points, and that the race was exactly tied (Blumenthal 2017). Reporting on polling data with uncertainty goes against the grain of what many journalists do today but the New York Times and SurveyMonkey, as well as other outlets like FiveThirtyEight, serve as examples that this type of reporting is possible and promotes rich discussion. It is important that journalists and pollsters continue to engage in this type of reporting, even if it is hard or makes for less catchy headlines.

The results of my analysis are instructive for future pollsters in two different ways. First, they suggest specific approaches that tend to work better. For pollsters that do not have access to historical vote validation data, creating a voter likelihood composite index, like the Perry-Gallup index or my reformulation of it, can be an effective method to narrow down likely voters with a cutoff approach. And it is really easy to justify to a broad audience. For pollsters that have more data at their disposal, training random forests models on historical data and using the vote propensity scores these models compute for the current respondents can be used as an additional survey weight to achieve accurate estimates. A problem that pollsters will run into here is transparency: while this modeling approach is effective, it is also challenging to explain. Even though its estimates are consistently accurate, this approach may not be right for a pollster who cannot defend how and why it is used. The unique scope of my project suggests these two approaches should lead to better polling estimates across the board, which should, in turn, lead to increased legitimacy for the polling industry and more accurate public opinion data for campaigns and scholars to study in the future.

One important limitation is that, since my analysis relies on CCES data, which is drawn from the voting age population at large, this project cannot offer strong recommendations for pollsters that sample off voter registration lists or voter files. In their 2016 report, Pew restricts its sample to registered voters and also considers the impact of adding voter file vote history into likely voter models, both of which are more challenging to do with the CCES data. Likely voter models based on samples drawn from these frames will naturally be different. For instance, respondents sampled from the voter file predict their voting behavior more accurately than the CCES respondents do (Cohn 2018).

The second takeaway is that probabilistic judgments can be readily made from polling data, especially by using likely voter models to look at different turnout scenarios. Again, this is where this project differs from other research that has been conducted about likely voter models, including the Pew report that I draw from deeply to set up my analysis. Other studies, mirroring what too many pollsters in the field do, have created some type of hard turnout cutoff or categorization. In my results section, I include visualizations and discussion of how the likely voter models I look at perform when very few potential voters end up voting, when all potential voters end up voting, and many situations in between. What I do is not groundbreaking or fancy and is easy for pollsters and news outlets to adopt into their own reporting practices.

Figures 17 and 18 are examples of visualizations that pollsters or journalists could use to convey their polling data in a probabilistic manner. A 2016 Florida pollster could have produced Figure 17 using a random forests-based likely voter model that considers Perry-Gallup and demographic variables in conjunction with a cutoff approach. On average, as shown in Figure

16, this type of model should have overestimated Trump's support by a handful of points. At the time this survey was fielded, Trump had a nearly three-point lead and went on to win the state by 1.2 points. The cutoff approach indicates that Trump led by between 2.8 and 11 points under reasonable turnout. When the vote propensity scores from this model are used as weights, this approach gives Trump a 0.2-point lead. Figure 18, on the other hand, is something a national pollster could produce using the Perry-Gallup index approach. The most accurate version of this model (using the preferences of respondents who score a 5 or 6) underestimates Clinton's lead at the time by 1.2 points and underestimates her actual margin of victory by 0.4 points. This version best reflects the turnout rate in 2008 and 2012 and gets close to the 2016 turnout rate of 54.7 percent (McDonald 2016).

[FIGURE 17 ABOUT HERE]

[FIGURE 18 ABOUT HERE]

One lingering issue with probabilistic forecasting is that these types of predictions may confuse people. This issue reemerged recently in light of Westwood et al.'s (2018) finding that probabilistic election predictions – those that give each candidate a win percentage rather than estimate each one's vote share – lead people to be overly confident that the leading candidate will win. Furthermore, people presented with win-probabilities (but not vote share estimates) are less likely to vote in a game that simulates an election (Westwood et al. 2018). Currently, polls' "effects on human behavior are not well understood" (Kennedy et al. 2016). If Westwood et al. are right, their conclusions have major implications for the use of probabilistic forecasting in election polling. Even if these models are the most accurate way to present polling information, what is their public value if people cannot understand them? And what if they work to actually demobilize potential voters? Do polling forecasts have a civic obligation to, at the minimum, not discourage people from participating in the democratic process?

My project combines the best of both camps. On one hand, by incorporating a range of turnouts into estimates, I am better able to quantify the uncertainty inherent in these types of predictions than if I used point estimates. On the other hand, my estimates still use candidate vote shares, which are more easily understood, while providing a different type of probabilistic information than win-probabilities. My approach suggests what should happen when turnout is higher or lower; instead of reducing the polling data into an estimate of how likely a candidate is to win it expands the amount of information that can be gleaned from the data. Does a high turnout election benefit the Democratic candidate? If turnout is roughly what it was in the last election, will the Republican candidate win? These are questions that win-probabilities do not answer, but an approach that emphasizes the uncertainty inherent in defining likely voters does. This type of information may even have a mobilizing effect; if a reader thinks that a high turnout election benefits (or hurts) their candidate, they might be more inclined to go to the polls or to encourage their friends to vote.

The challenging part of studying elections and voting behavior, however, is that new data is constantly being generated and should be added to likely voter models. It will be important to regularly update this type of research with the inclusion of new data. Along these lines, I believe that the project of modeling likely voters would benefit from a Bayesian approach, which was outside of the scope of this project. The Bayesian method of setting a prior belief about the likelihood of an event occurring and updating that belief as more information becomes available maps nicely onto figuring out what the chances are that a survey respondent will vote in an upcoming election. A simple Bayesian analysis could establish priors using respondents' self-reported vote intentions (which are not necessarily the most accurate estimator of voting behavior) and update these priors using much of the same information I look at here.

It should also be noted that this research focuses specifically on using likely voter models to generate accurate vote share estimates. I consider this the most useful criteria to use for judging, since this is how pundits and the public typically assess polls. But analyses of poll performance can just as legitimately focus on predicting turnout or predicting the demographic composition of the electorate (Kenney et al. 2016). Indeed, if likely voter models are to be theoretically grounded, it is important that they do a good job of capturing more than just the result of the race. This project uses the sample's demographics and considers potential turnout rates as tools to achieve more accurate vote share estimates, but future research can and should focus on how well likely voter models predict turnout and electoral composition. On the other hand, future research should seek to identify which demographic variables I include in my logistic regression and random forests models are the most important for prediction to reduce the data-gathering burden on pollsters.

10. Figures

Table 1. Validated vote by self-reported voting intentions for 2016 CCES

Vote intention	**Voted**	**Did not vote**
Yes, definitely	62%	38%
Probably	28%	72%
I already voted (early or absentee)	67%	33%
No	5%	95%
Undecided	15%	85%

Table 2. Previous findings on what predicts who votes and who misreports

	Voting	**Misreporting**
Gender	Ansolabehere and Hersh, 2012; Leighley and Nagler, 2013	
Age	Verba and Nie, 1972; Wolfinger and Rosenstone, 1980; Blais, 2006; Ansolabehere and Hersh, 2012; Leighley and Nagler, 2013	Rogers and Aida, 2014; Pew Research Center, 2000
Race	Verba and Nie, 1972; Ansolabehere and Hersh, 2012	Bernstein et al., 2001; Rogers and Aida, 2014
Education	Verba and Nie, 1972; Wolfinger and Rosenstone, 1980; Blais, 2006; Ansolabehere and Hersh, 2012	Ansolabehere and Hersh, 2012; Rogers and Aida, 2014; Pew Research Center, 2000
Income	Ansolabehere and Hersh, 2012; Leighley and Nagler, 2013	Ansolabehere and Hersh, 2012
Partisan/ideological strength	Verba and Nie, 1972; Ansolabehere and Hersh, 2012	Bernstein et al., 2001; Ansolabehere and Hersh, 2012; Rogers and Aida, 2014
Religiosity	Ansolabehere and Hersh, 2012	Bernstein et al., 2001; Ansolabehere and Hersh, 2012
Composition of district	Verba and Nie, 1972	Bernstein et al., 2001
Marital status	Wolfinger and Rosenstone, 1980); Ansolabehere and Hersh, 2012	
Residential mobility	Ansolabehere and Hersh, 2012	Ansolabehere and Hersh, 2012
Political interest/activism	Verba and Nie, 1972; Ansolabehere and Hersh, 2012	Bernstein et al., 2001; Ansolabehere and Hersh, 2012

Table 3. CCES sample sizes, 2008-2016

Year	**Sample size**
2008	32,800
2010	55,400
2012	54,535
2014	56,200
2016	64,600

Figure 1. Election predictions for 2016 general election by vote intent

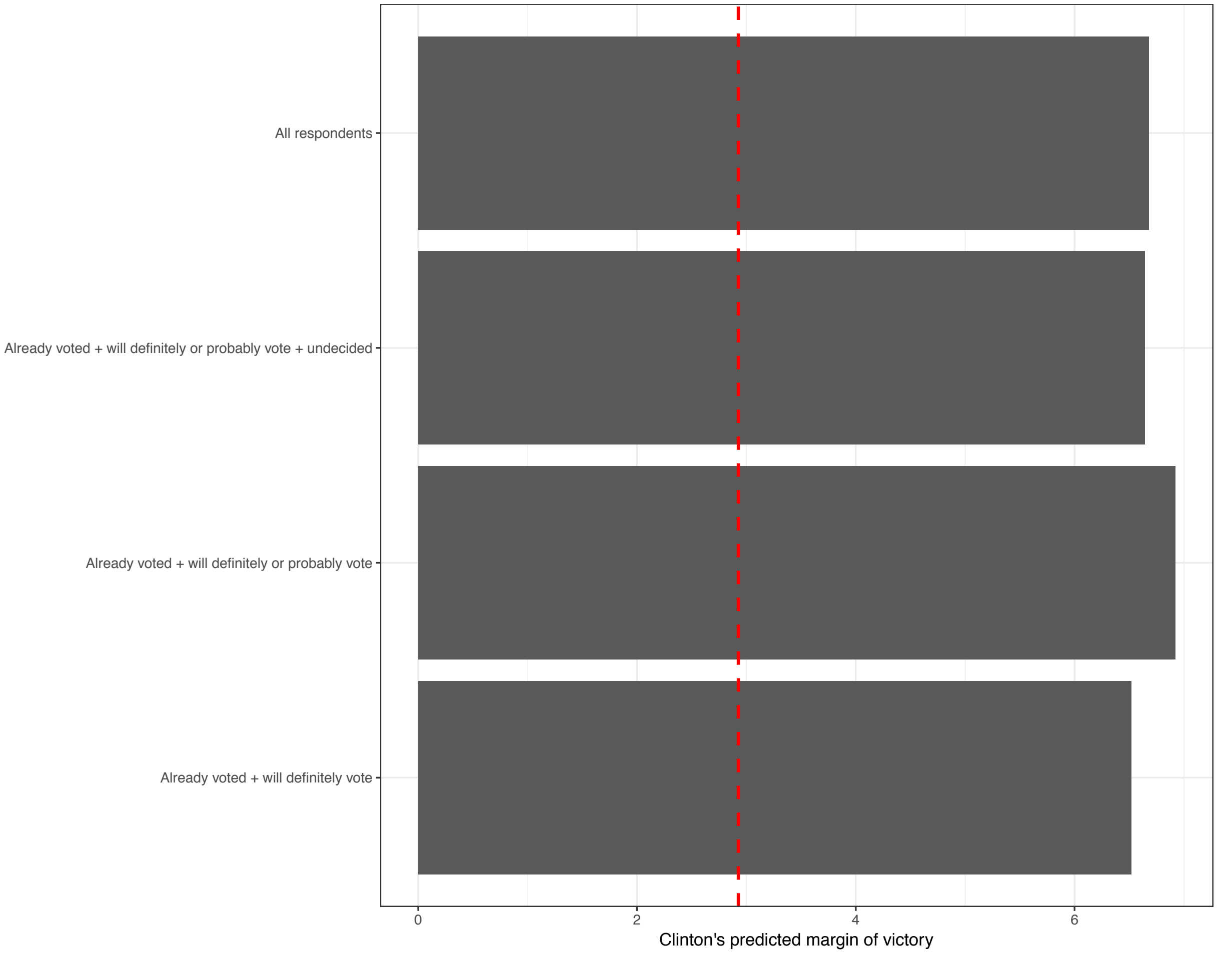


Figure 2. Election predictions for state-level 2016 general election by vote intent

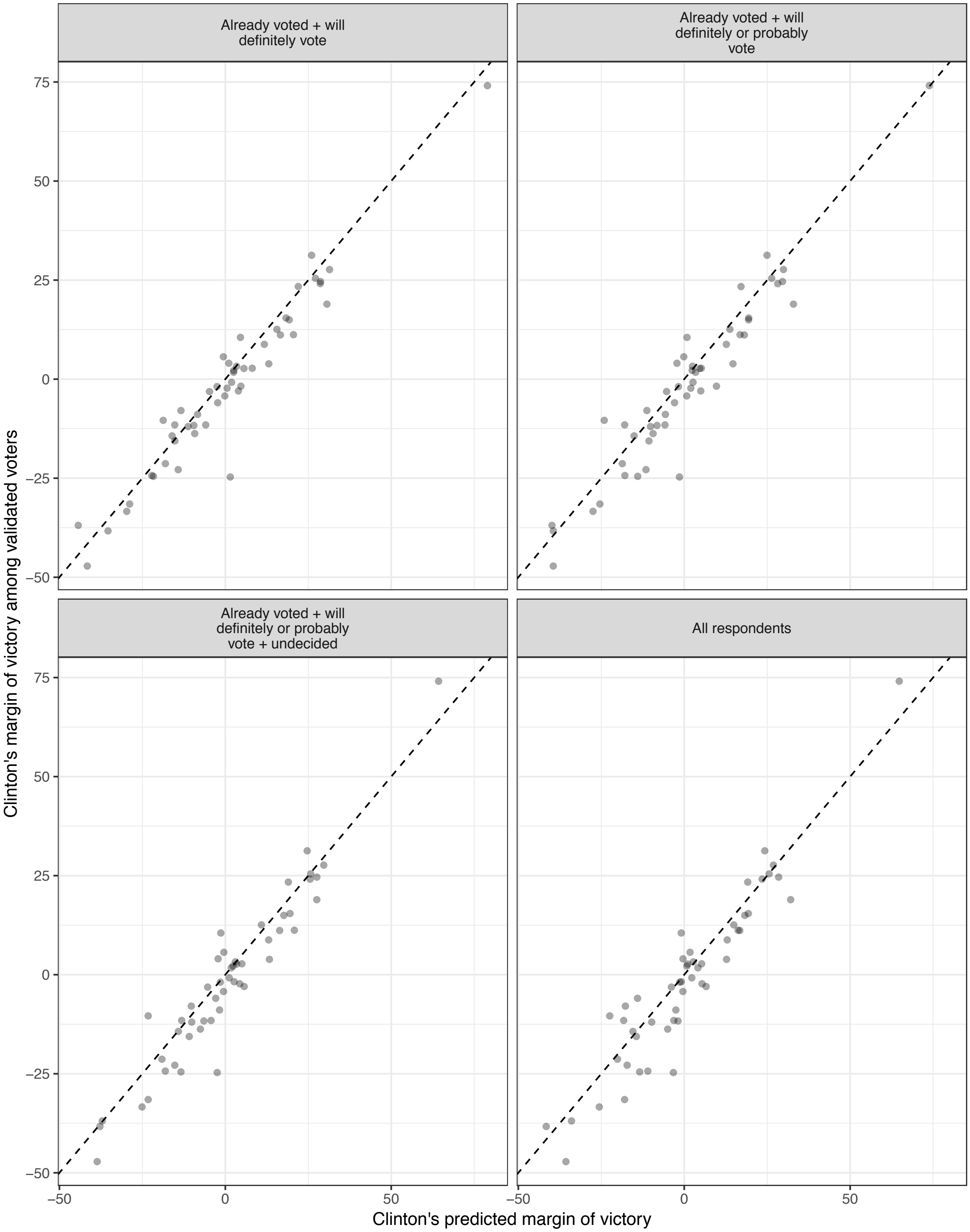


Figure 3. Election predictions for 2016 general election by vote intent and vote history

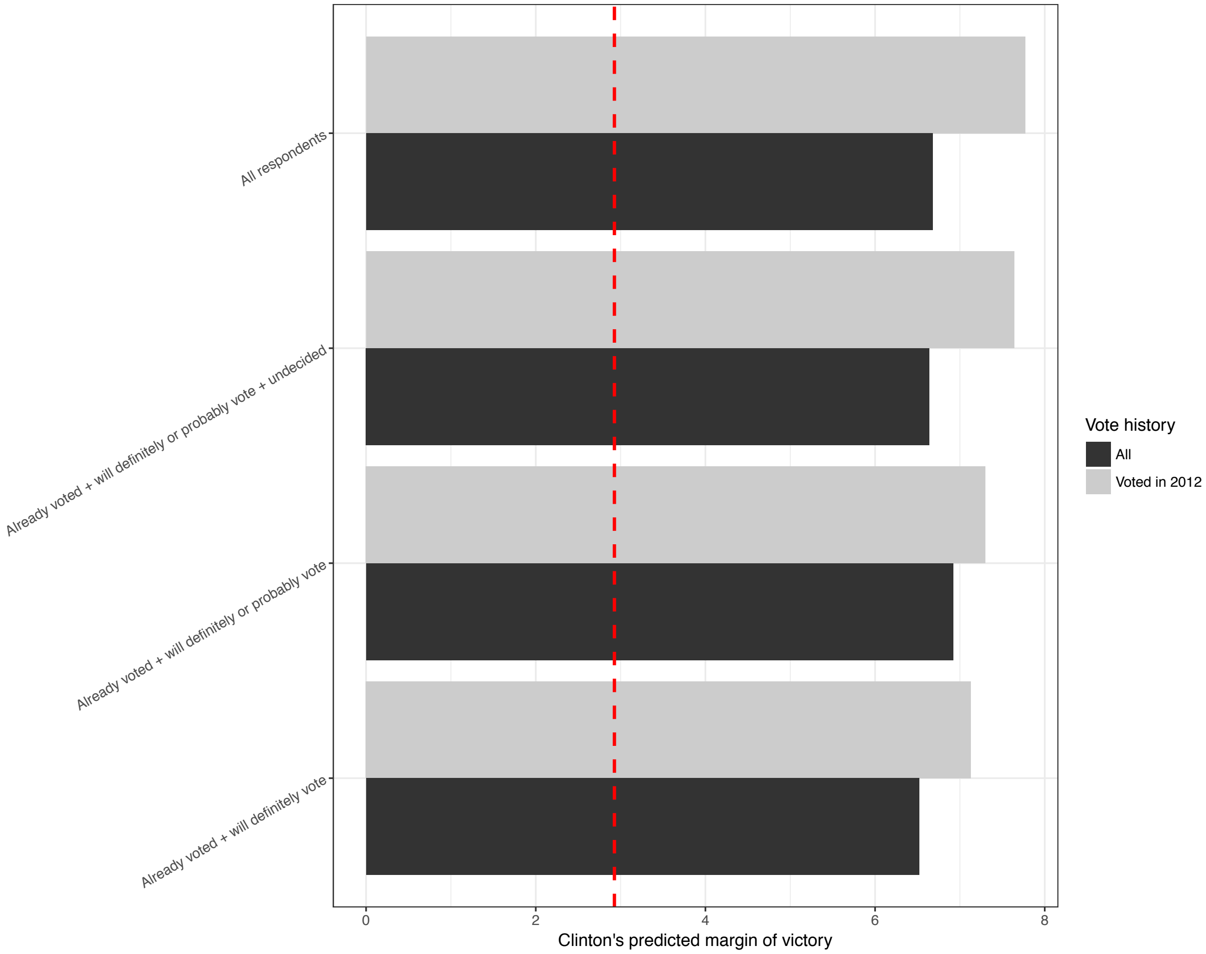


Figure 4. Election predictions for state-level 2016 general election by vote intent for self-reported 2012 general election voters

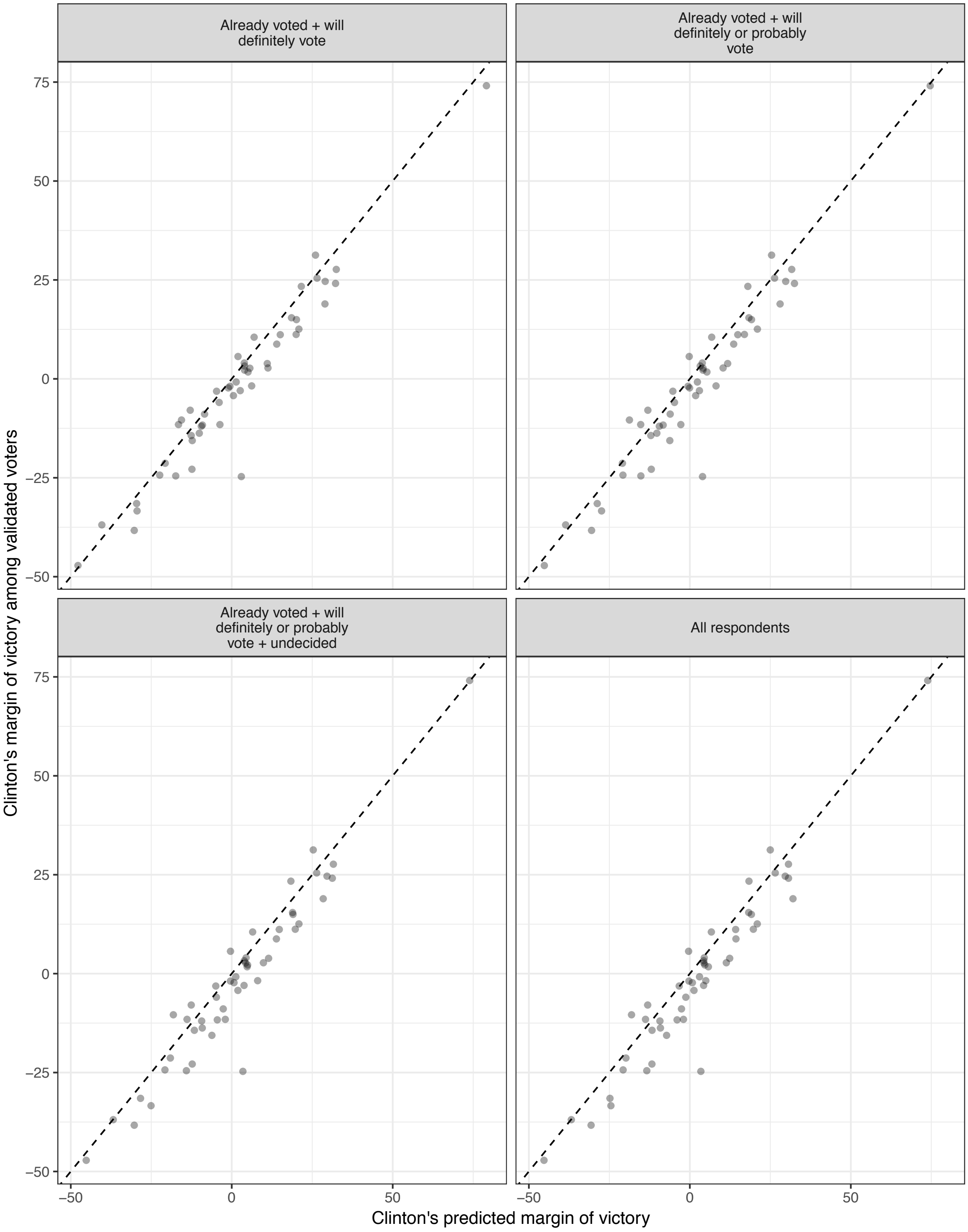


Figure 5. Election predictions for 2016 general election by likely voter index scores using a variation of the Perry-Gallup index

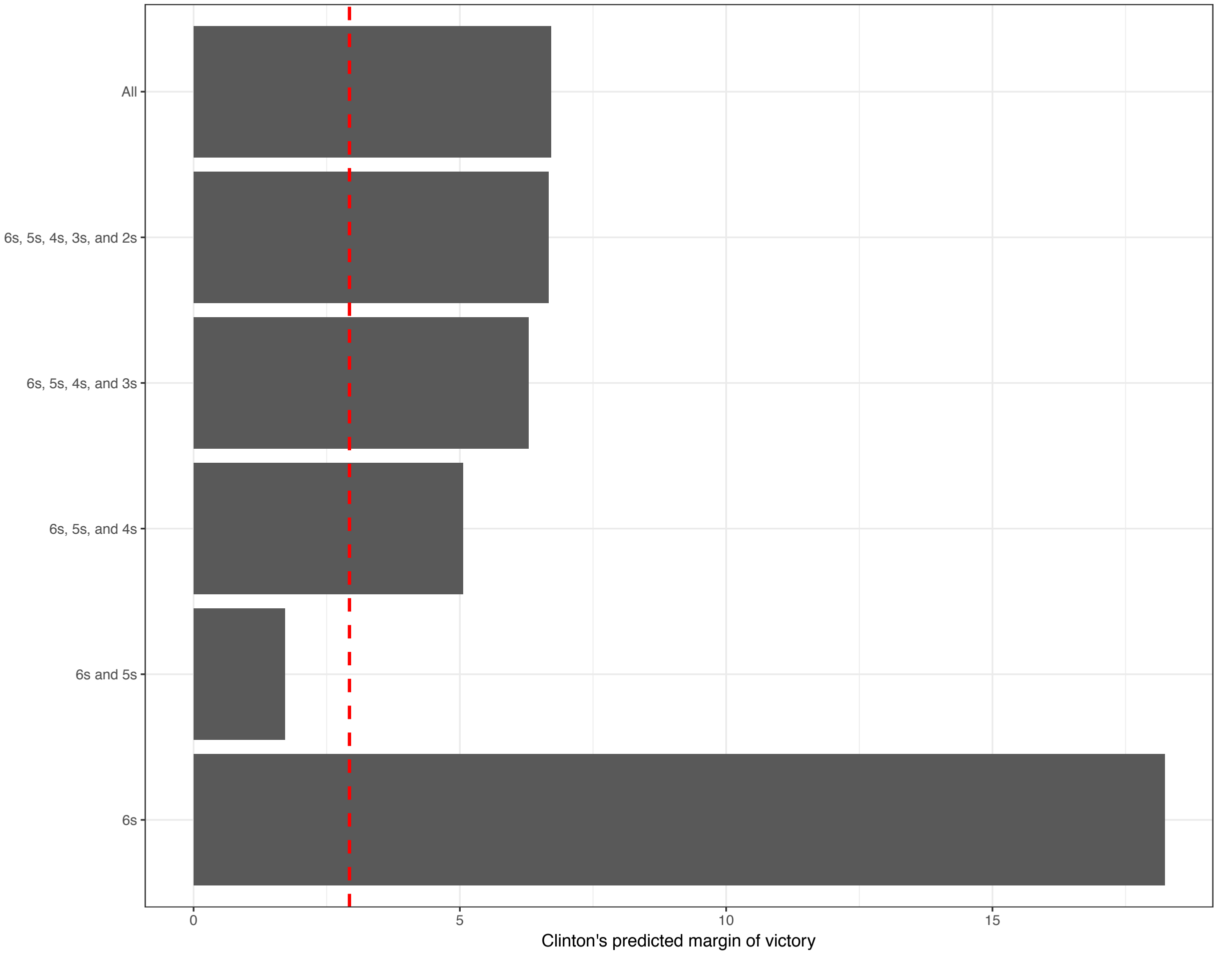


Figure 6. Election predictions for state-level 2016 general election by likely voter index scores using a variation of the Perry-Gallup index

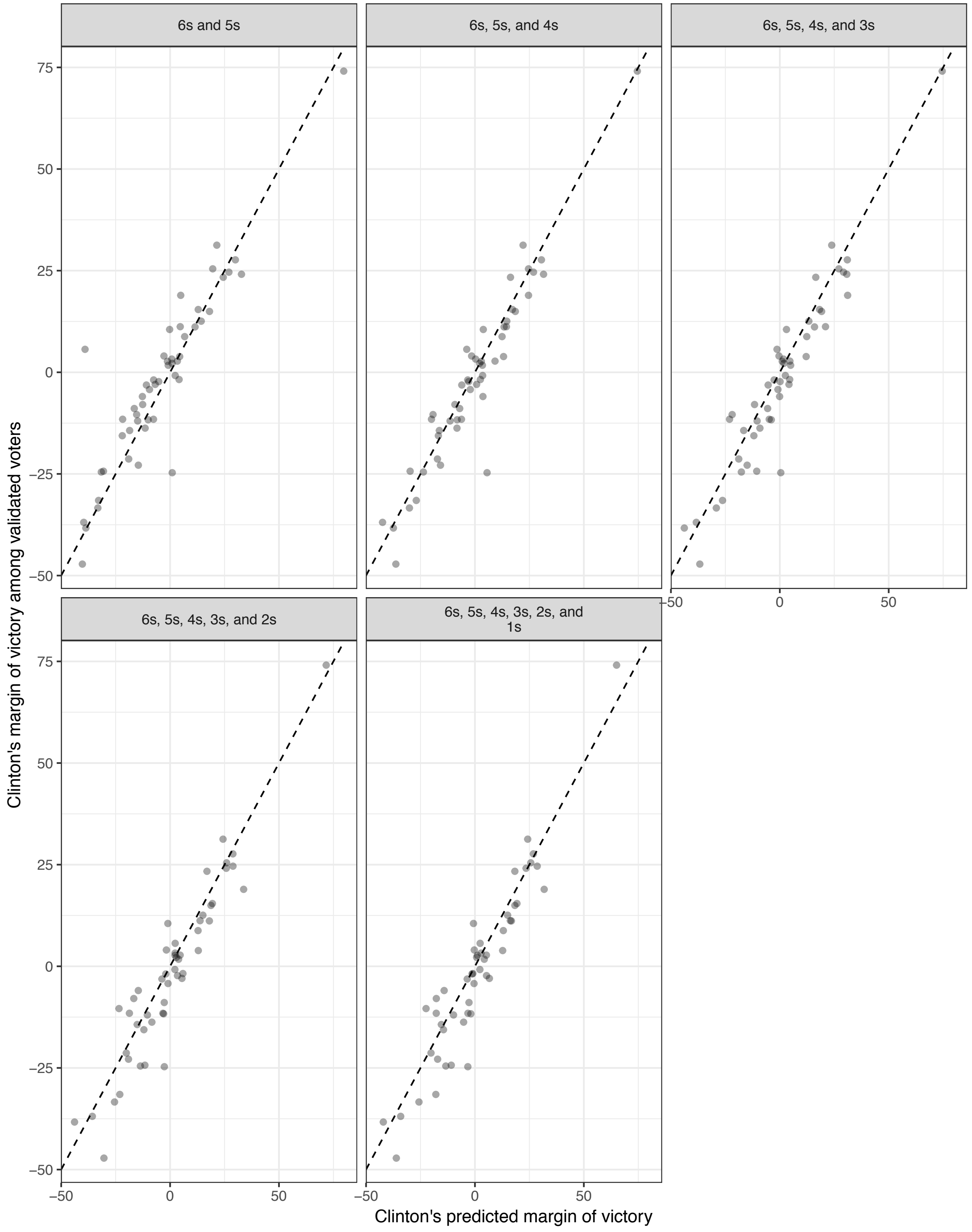
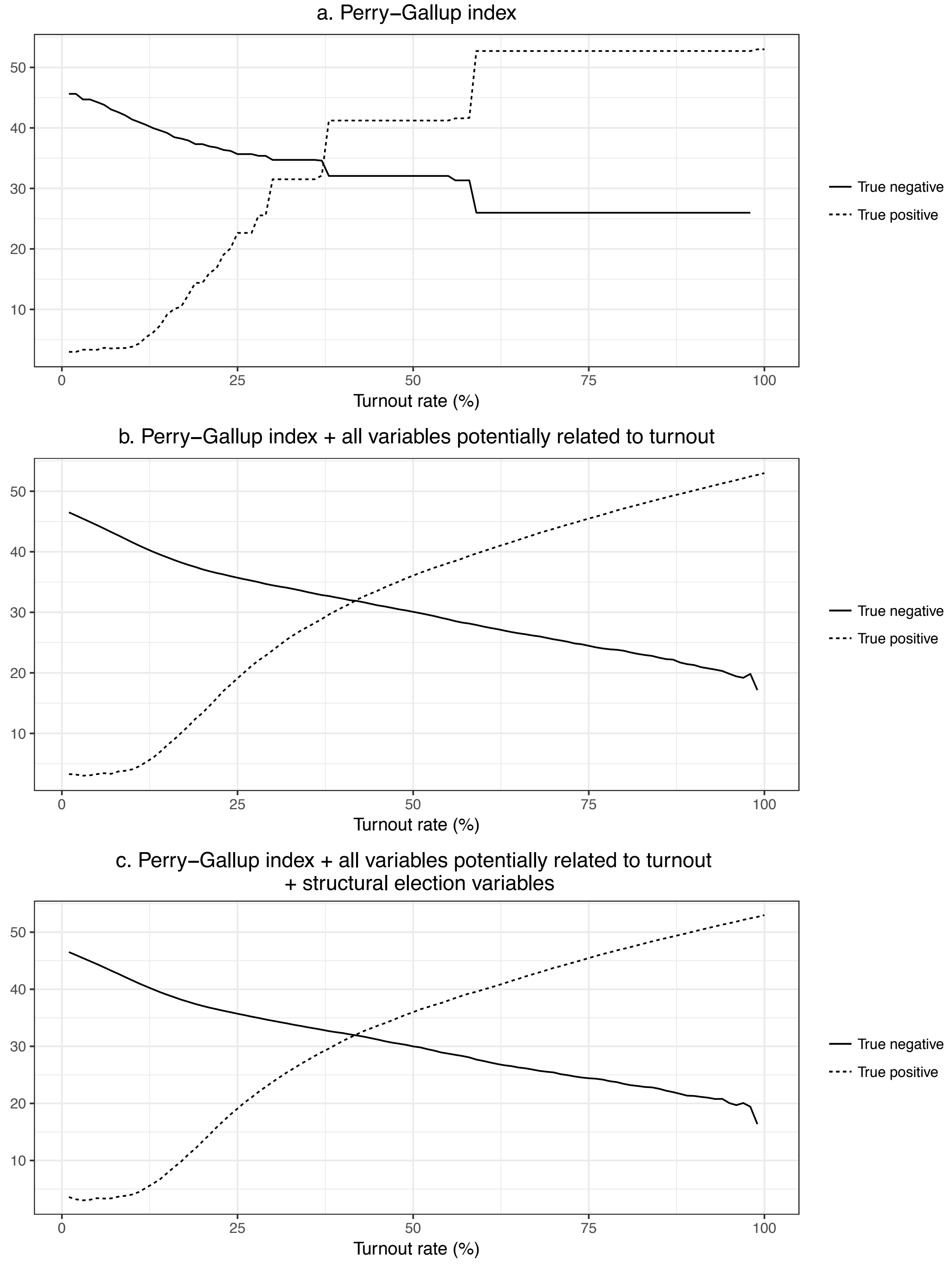


Figure 7. True positive and true negative rates by turnout with logistic regression approach

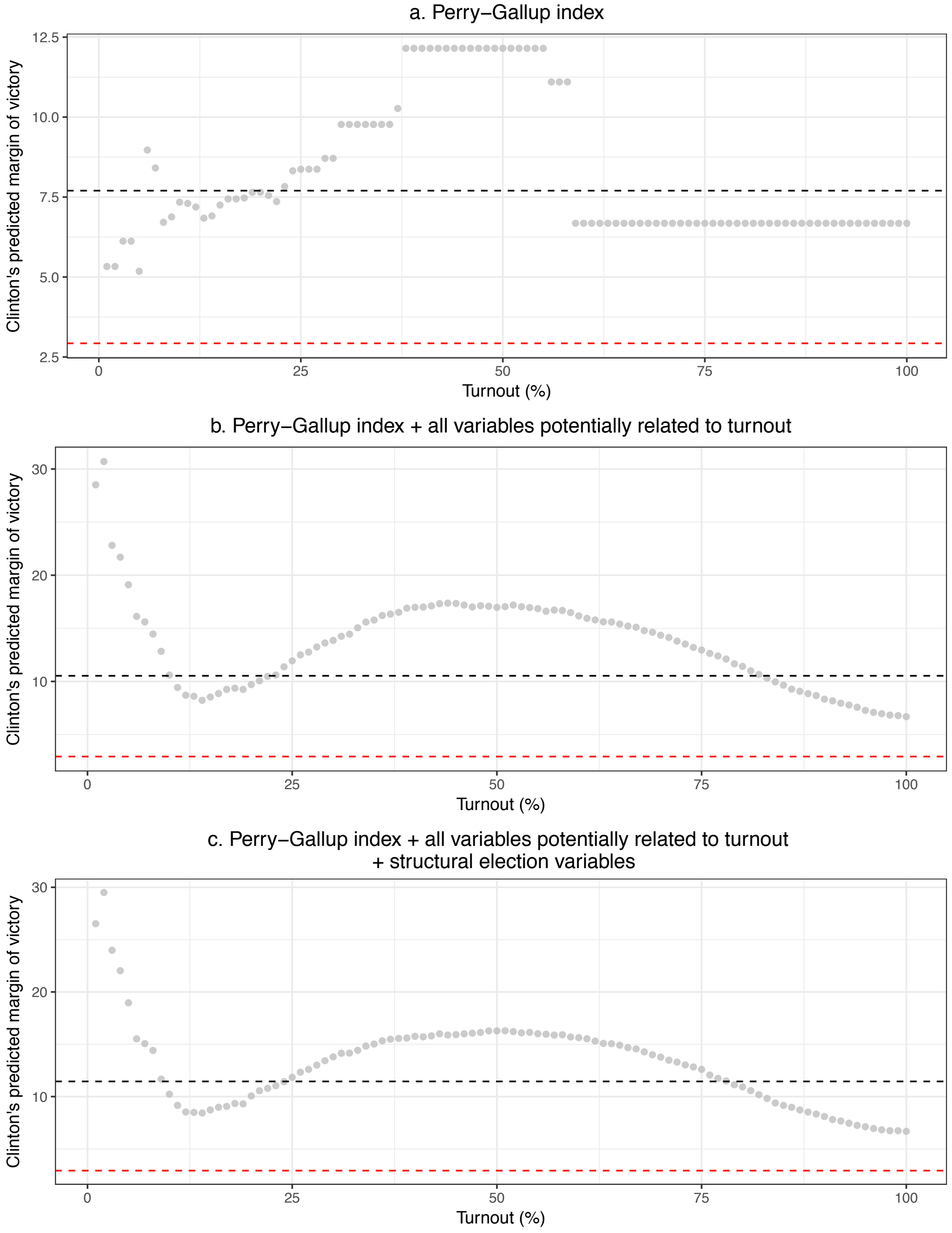


a. Perry−Gallup index

b. Perry−Gallup index + all variables potentially related to turnout

c. Perry−Gallup index + all variables potentially related to turnout + structural election variables

Figure 8. Election predictions for 2016 general election by turnout with logistic regression approach



The red dashed line is the margin among validated voters in the 2016 CCES. The black dashed line is the vote-propensity-weighted margin among all 2016 CCES respondents.

Figure 9. Error between state-level predicted and actual margin of victory for Clinton in 2016 general election by turnout using logistic regression approach

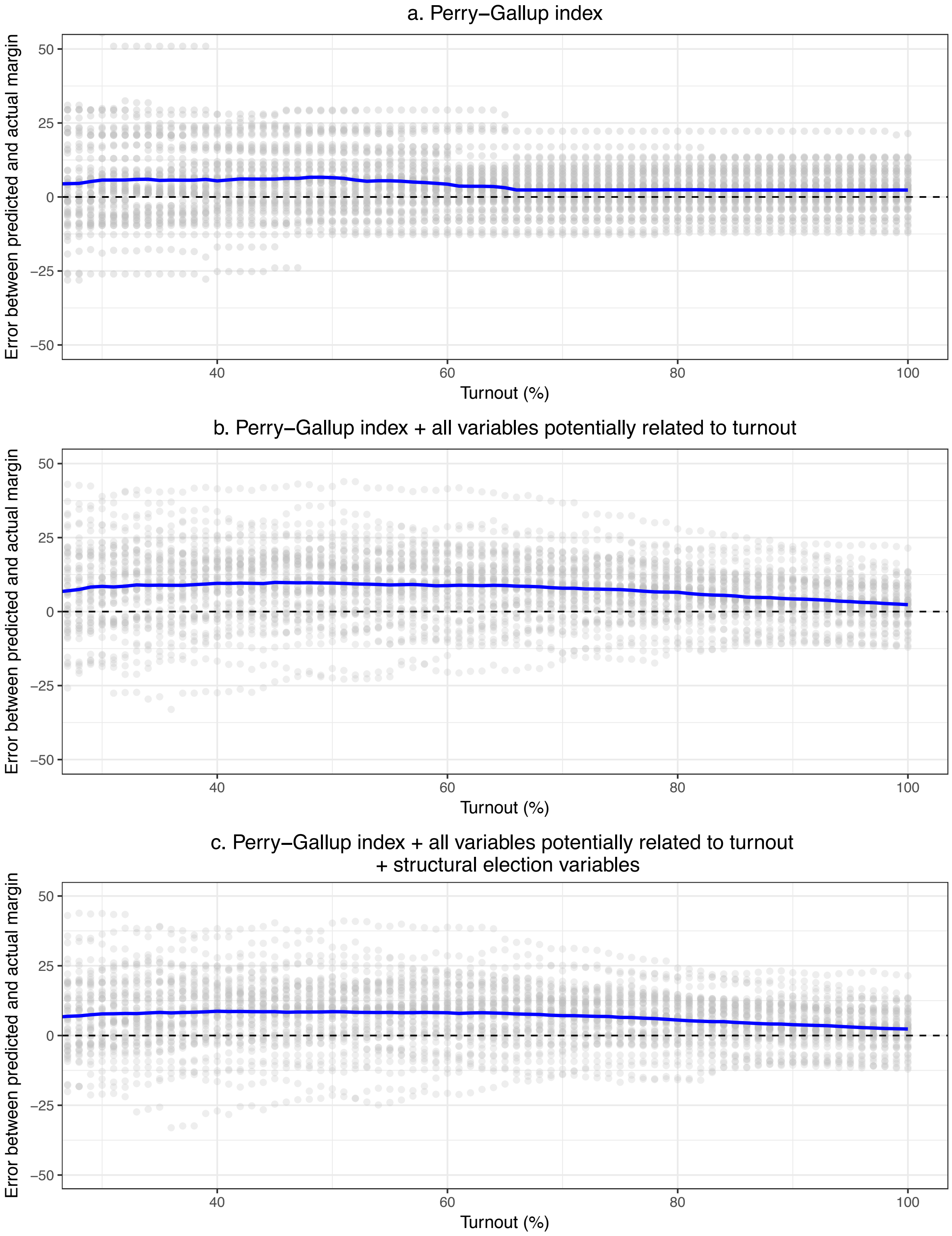


Figure 10. Election predictions for state-level 2016 general election using logistic regression-based vote propensity weight

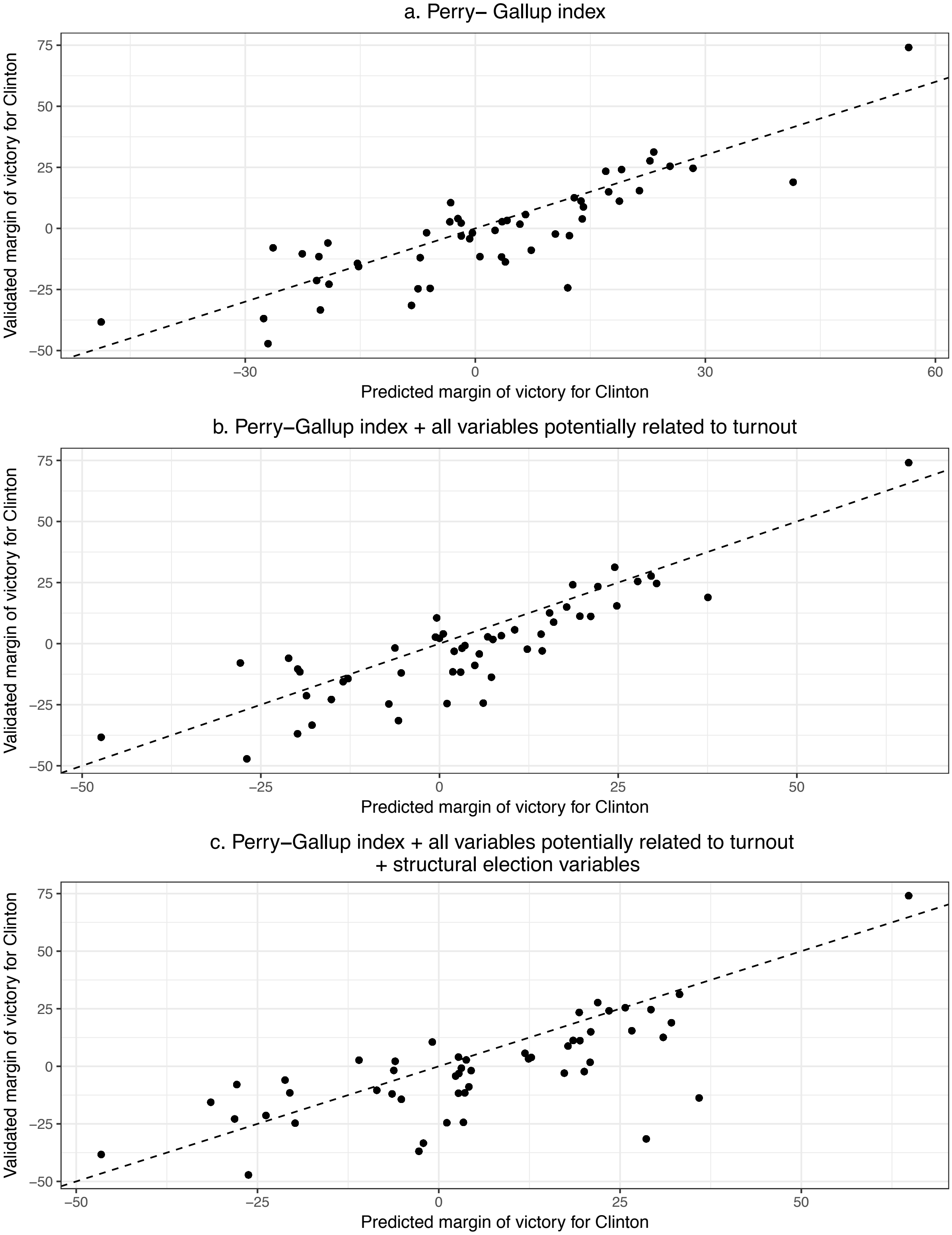
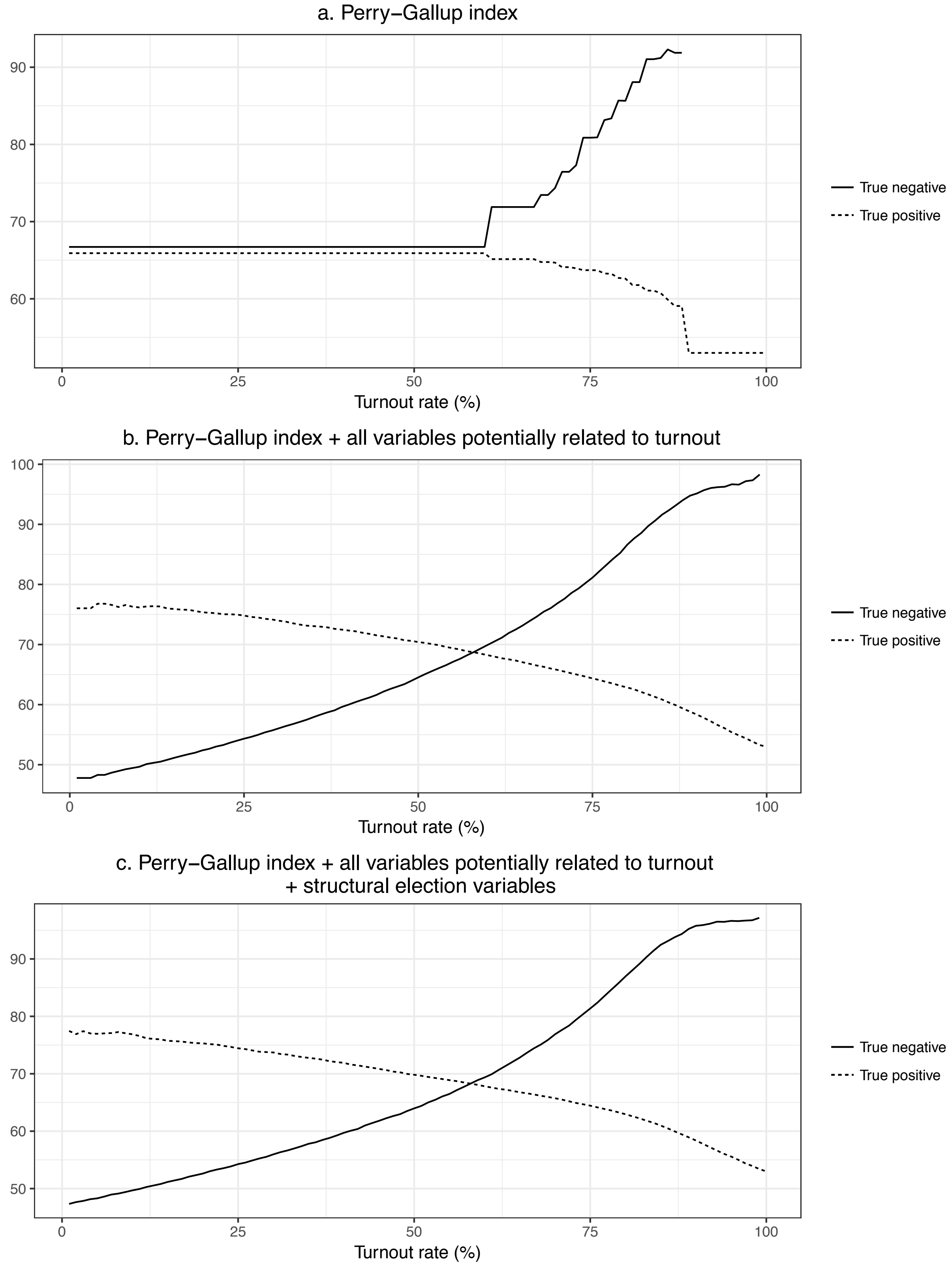


Figure 11. True positive and true negative rates by turnout with random forest approach

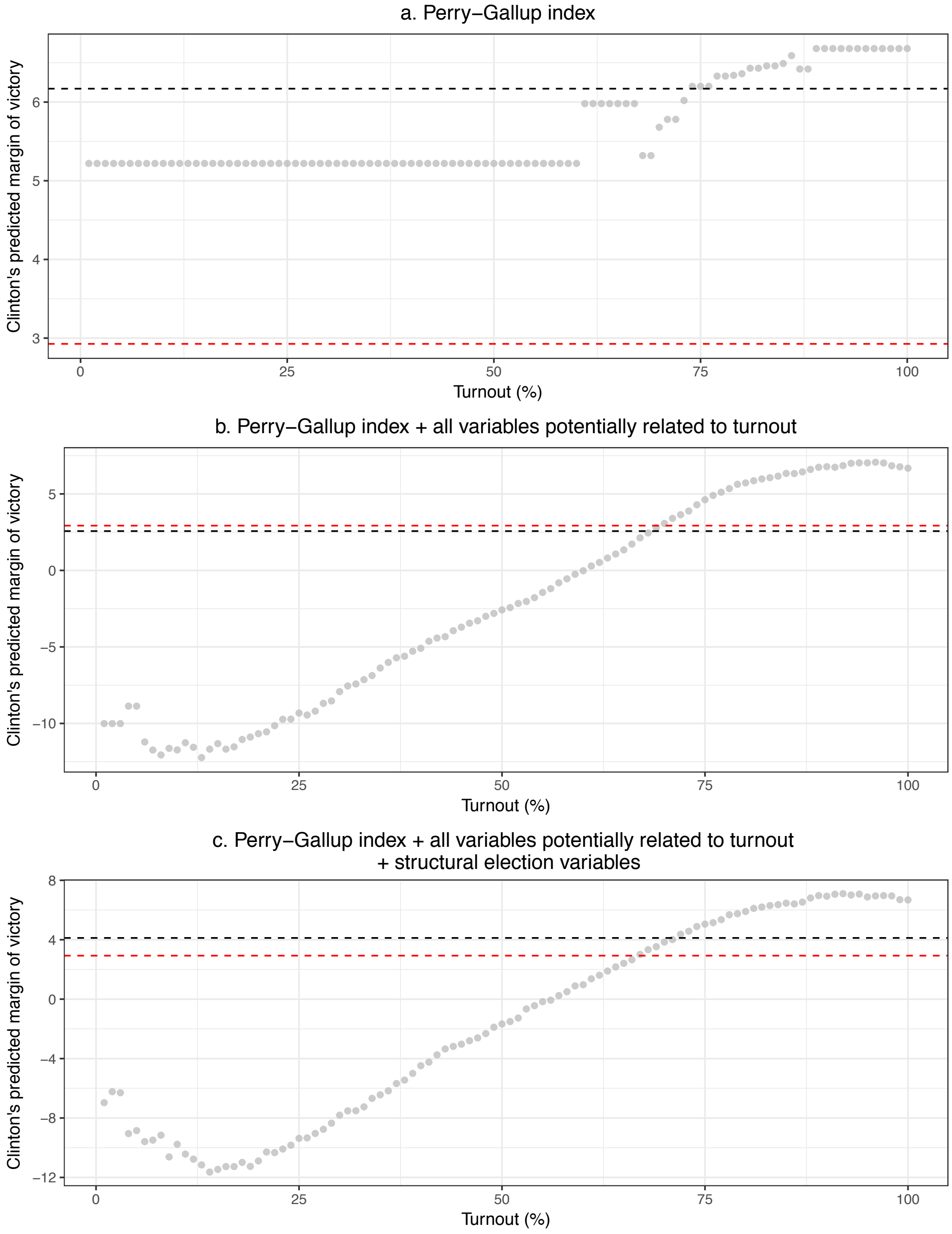


a. Perry–Gallup index

b. Perry–Gallup index + all variables potentially related to turnout

c. Perry–Gallup index + all variables potentially related to turnout
+ structural election variables

Figure 12. Election predictions for 2016 general election by turnout with random forest approach



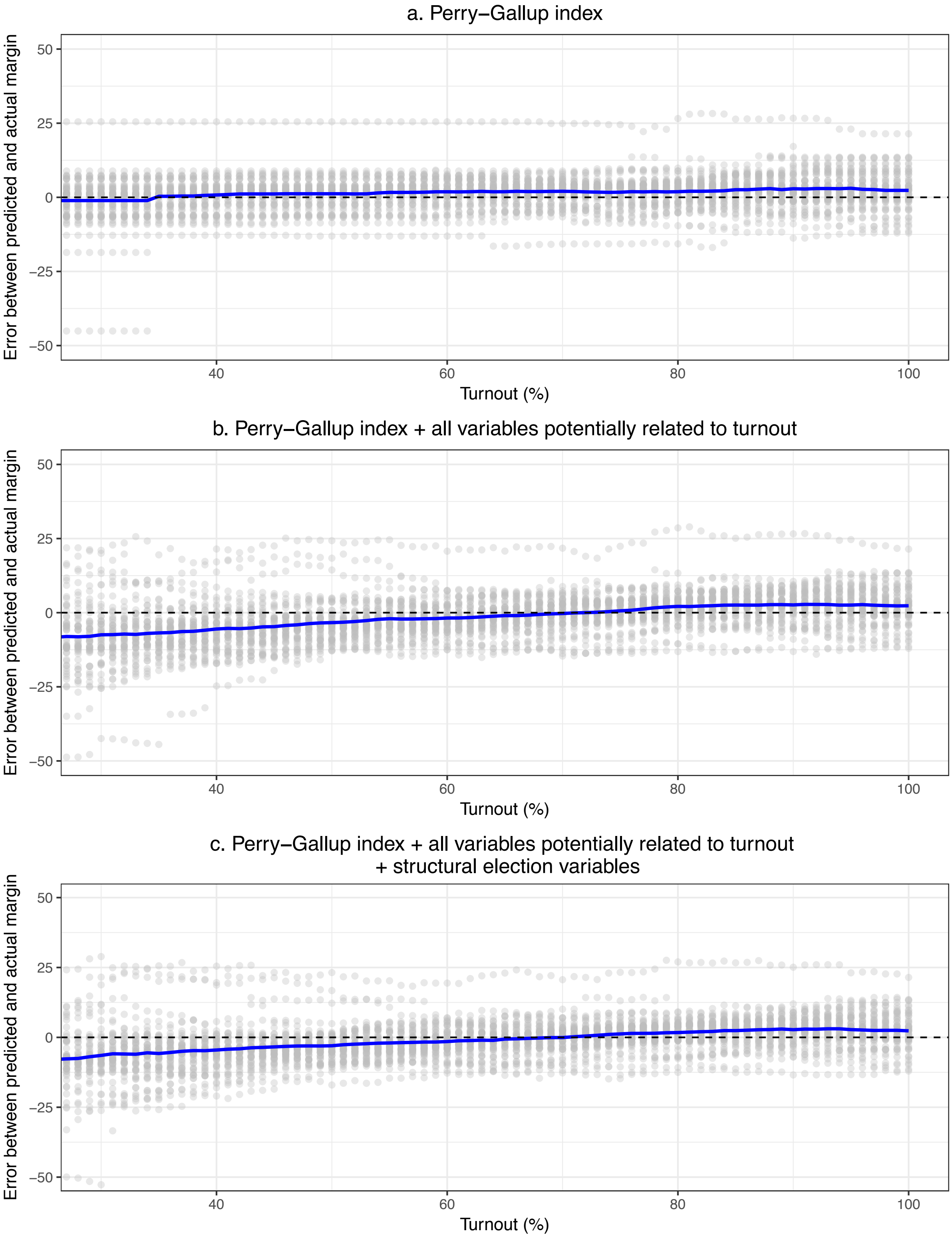
a. Perry–Gallup index

b. Perry–Gallup index + all variables potentially related to turnout

c. Perry–Gallup index + all variables potentially related to turnout + structural election variables

The red dashed line is the margin among validated voters in the 2016 CCES. The black dashed line is the vote-propensity-weighted margin among all 2016 CCES respondents.

Figure 13. Error between state-level predicted and actual margin of victory for Clinton in 2016 general election by turnout using random forest approach

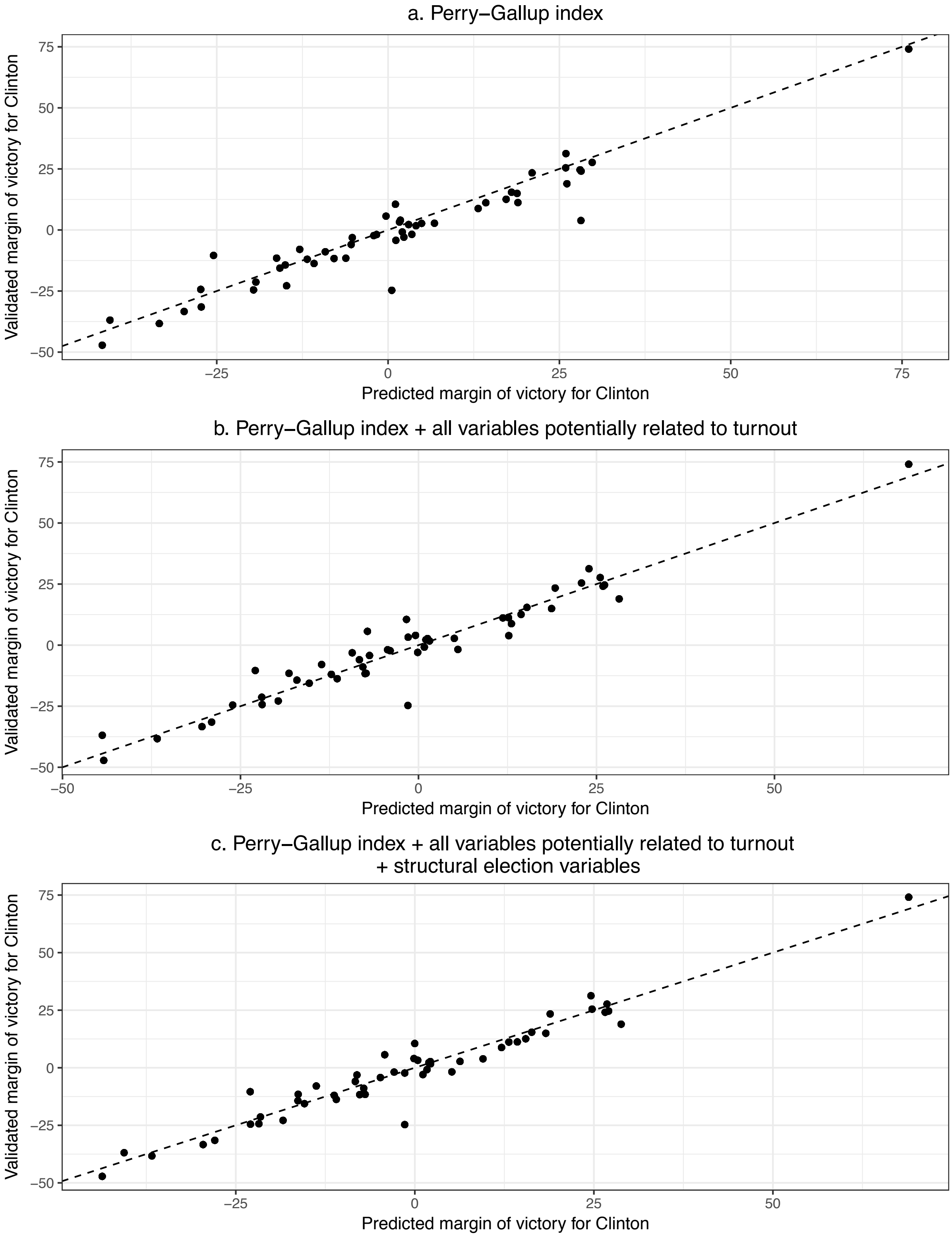


a. Perry–Gallup index

b. Perry–Gallup index + all variables potentially related to turnout

c. Perry–Gallup index + all variables potentially related to turnout + structural election variables

Figure 14. Election predictions for state-level 2016 general election using random forest-based vote propensity weight



a. Perry−Gallup index

b. Perry−Gallup index + all variables potentially related to turnout

c. Perry−Gallup index + all variables potentially related to turnout + structural election variables

Figure 15. Error for national-level models

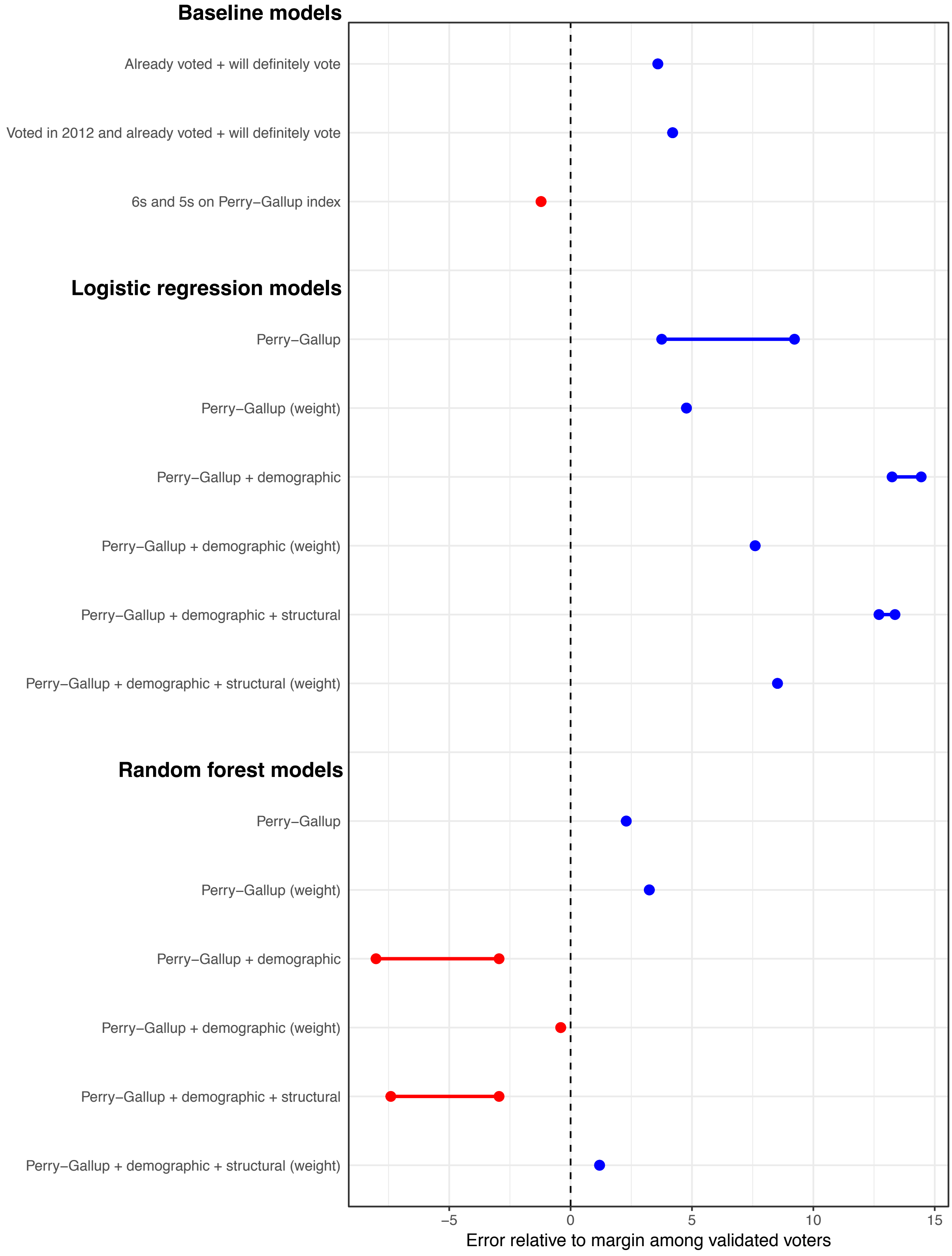


Figure 16. Error for state-level models

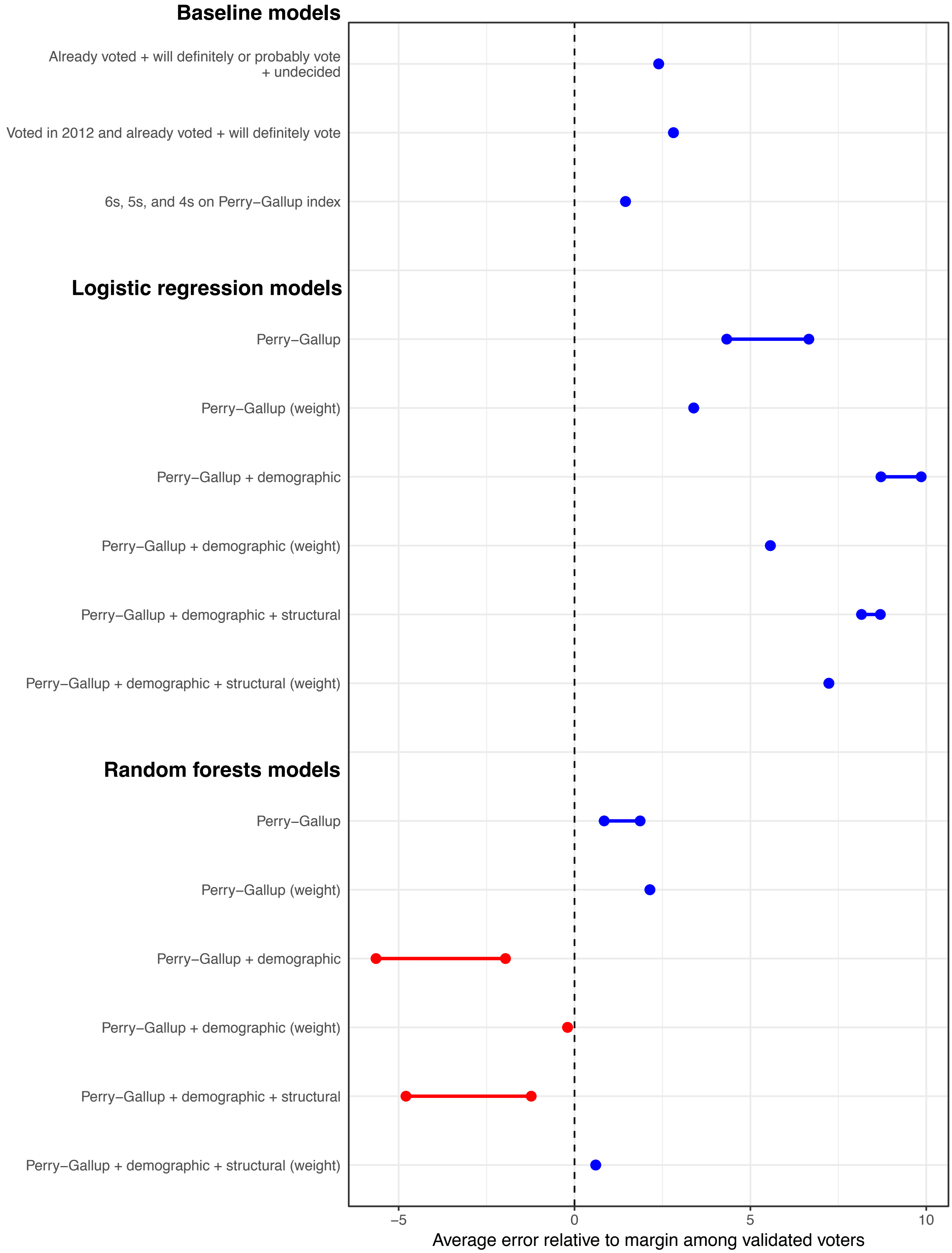


Figure 17. Mock-up visualization from a 2016 Florida pollster

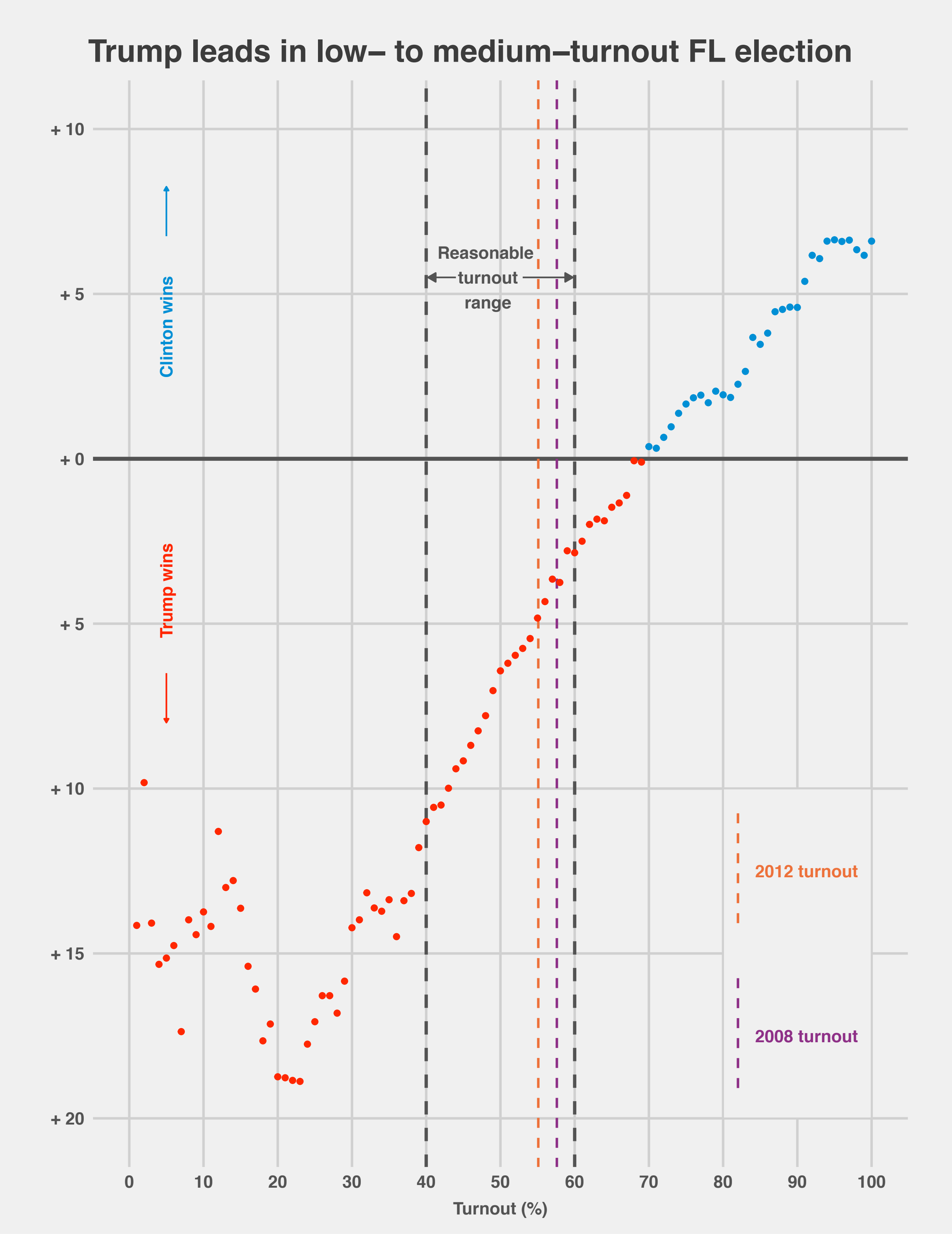
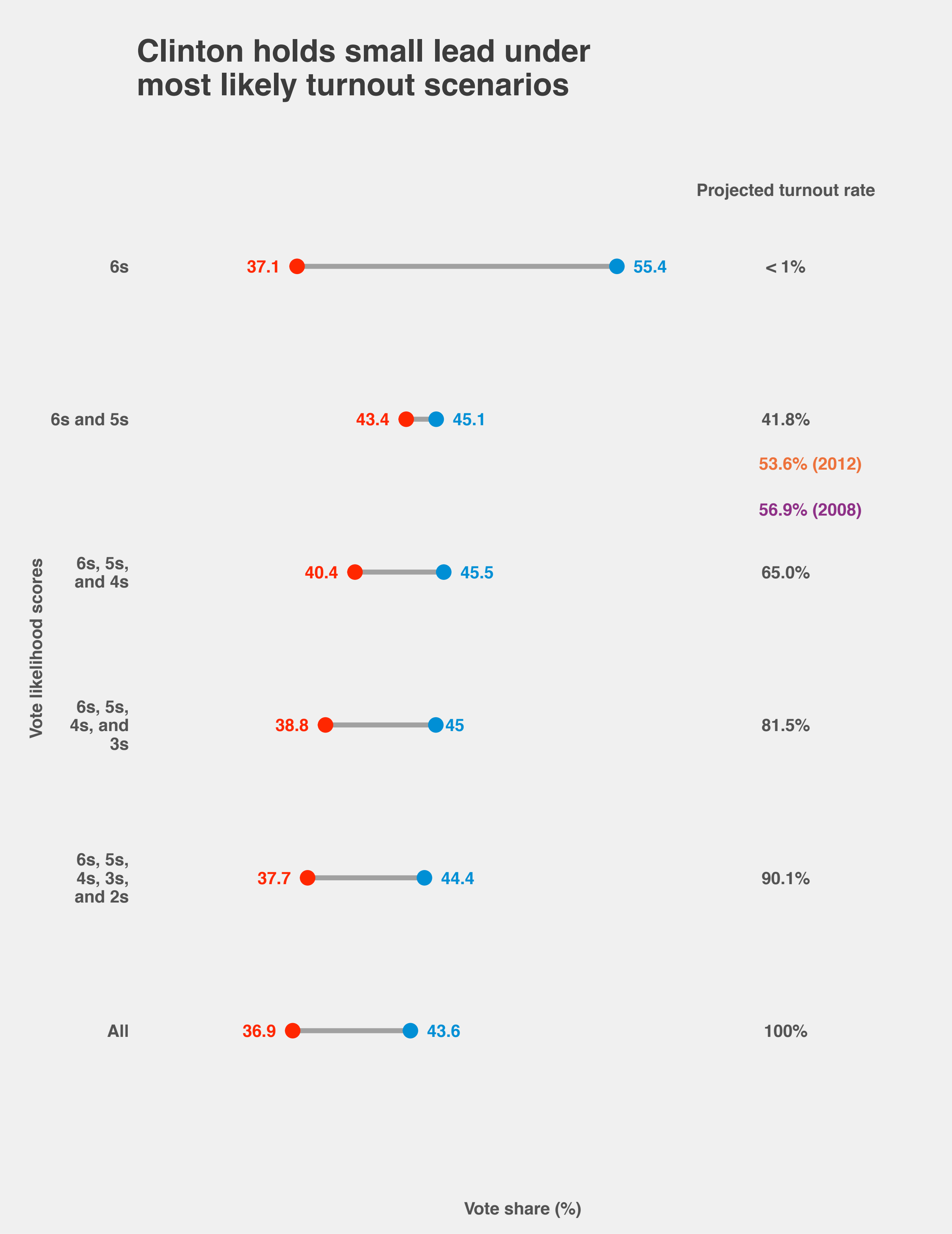


Figure 18. Mock-up visualization from a 2016 national pollster



11. References

Abramowitz, A. (2012). Forecasting in a polarized era the time for change model and the 2012 presidential election. *PS: Political Science & Politics,* (4), 618.

Abramowitz, A. I. (2008). Forecasting the 2008 presidential election with the time-for-change model. *PS: Political Science and Politics,* (4), 691.

Ansolabehere, S. COOPERATIVE CONGRESSIONAL ELECTION STUDY, 2008: COMMON CONTENT. [computer file] release 4: July 15, 2011. Cambridge, MA: Harvard University [producer].

Ansolabehere, S. COOPERATIVE CONGRESSIONAL ELECTION STUDY, 2010: COMMON CONTENT. [computer file] release 1: August 1, 2012. Cambridge, MA: Harvard University [producer].

Ansolabehere, S. COOPERATIVE CONGRESSIONAL ELECTION STUDY, 2012: COMMON CONTENT. [computer file] release 1: April 15, 2013. Cambridge, MA: Harvard University [producer].

Ansolabehere, S. COOPERATIVE CONGRESSIONAL ELECTION STUDY, 2014: COMMON CONTENT. [computer file] release 1: November 13, 2016. Cambridge, MA: Harvard University [producer].

Ansolabehere, S., & Hersh, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Political Analysis, 20*(4), 437-459.

Ansolabehere, S., & Schaffner, B. F. COOPERATIVE CONGRESSIONAL ELECTION STUDY, 2016: COMMON CONTENT [computer file] release 2: August 4, 2017. Cambridge, MA: Harvard University [producer].

Berent, M., Krosnick, J. A., & Lupia, A. (2011). *The quality of government records and Over-estimation of registration and turnout in surveys: Lessons from the 2008 ANES panel Study's registration and turnout validation exercises.* Working Paper no. nes012554. Ann Arbor, MI, and Palo Alto, CA: American National Election Studies.

Bernstein, R., Chadha, A., & Montjoy, R. (2001). Overreporting voting: Why it happens and why it matters. *Public Opinion Quarterly, 65*(1), 22-44.

Blais, A. (2006). What affects voter turnout? *Annu.Rev.Polit.Sci., 9*, 111-125.

Blumenthal, M. (2017). *Alabama senate race: A poll without a prediction.* Retrieved April 9, 2018, from https://www.surveymonkey.com/curiosity/alabama-senate-race-a-poll-without-a-prediction/.

Bolger, G., David, D. W., Franklin, C., Groves, R. M., Lavrakas, P. J., Mellman, M. S., et al. (2008). *Report of the AAPOR ad hoc committee on the 2008 presidential primary polling*American Association of Public Opinion Research.

Christensen, W., Christensen, W. F., & Florence, L. W. (2008). *Predicting presidential and other multistage election outcomes using state-level pre-election polls*.

Clement, S. (2016, January 7, 2016). Why the 'likely voter' is the holy grail of polling. *The Washington Post.*

Cohn, N. (2016, We gave four good pollsters the same raw data. they had four different results. . *New York Times: The Upshot.*

Cohn, N. (2018, April 3, 2018, 11:45 a.m.). "Outtake for twitter: Accuracy of self-reported likely vote intention in Upshot/Siena polls. includes all 7 polls we've conducted so far." (Twitter). Message posted to https://twitter.com/Nate_Cohn/status/981241232746860544.

Crespi, I. (1988). *Pre-election polling: Sources of accuracy and error.* Russell Sage Foundation.

Elon Poll. (2016). *ELON POLL: Presidential, governor's races too close to call in N.C., with voters worried about next president's decisions.* Retrieved April 26, 2018, from https://www.elon.edu/e/elon-poll/poll-archive/110116.

Erikson, R. S., Panagopoulos, C., & Wlezien, C. (2004). Likely (and unlikely) voters and the assessment of campaign dynamics. *Public Opinion Quarterly, 68*(4), 588-601.

Freedman, P., & Goldstein, K. (1996). Building a probable electorate from preelection polls: A two-stage approach. *Public Opinion Quarterly, 60*(4), 574-587.

Gallup Poll. *Presidential approval ratings -- barack obama.* Retrieved February 6, 2018, 2018, from http://news.gallup.com/poll/116479/barack-obama-presidential-job-approval.aspx.

Gallup Poll. *Presidential approval ratings -- george W. bush.* Retrieved February 6, 2018, 2018, from http://news.gallup.com/poll/116500/presidential-approval-ratings-george-bush.aspx.

Gelman, A., & King, G. (1993). Why are american presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science, 23*(4), 409-451.

Highton, B., McGhee, E., & Sides, J. (2014). Election fundamentals and polls favor the republicans. *PS-Political Science and Politics, 47*(4).

Hillygus, D. S. (2011). The evolution of election polling in the united states. *Public Opinion Quarterly, 75*(5), 962-981.

Hoek, J., & Gendall, P. (1997). Factors affecting political poll accuracy: An analysis of undecided respondents. *Marketing Bulletin-Department of Marketing Massey University, 8*, 1-14.

Holbrook, T. M., & DeSart, J. A. (1999). Using state polls to forecast presidential election outcomes in the american states. *International Journal of Forecasting, 15*(2), 137-142.

Jackman, N. (2016). Election forecasting in the media. In A. Therriault (Ed.), *Data and Democracy: How political data science is shaping the 2016 elections* (First ed., pp. 39). Sebastopol, CA: O'Reilly Media.

Keith, B., Magleby, D., Nelson, C., Orr, E., Westlye, M., & Wolfinger, R. (1986). The partisan affinities of independent 'Leaners'. *British Journal of Political Science, 16*(2), 155-185.

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. d., Durand, C., Franklin, C., et al. (2016). *An evaluation of the 2016 election polls in the U.S.* American Association of Public Opinion Research.

Kiley, J., & Dimock, M. (2009). *Understanding likely voters* (Methodological Note). Pew Research Center.

Leighley, J. E., & Nagler, J. (2013). *Who Votes Now?: Demographics, Issues, Inequality, and Turnout in the United States.* Princeton University Press.

Linzer, D. A. (2013). Dynamic bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association, 108*(501), 124-134.

Mannering, V. (2008). *GROSS DOMESTIC PRODUCT: Second quarter 2008 (advance) REVISED ESTIMATES: 2005 THROUGH FIRST QUARTER 2008.* No. BEA 08-34. Bureau of Economic Analysis.

Mataloni, L. (2012). *National income and product accounts gross domestic product, 2nd quarter 2012 (advance estimate); revised estimates: 2009 through first quarter 2012.* No. BEA 12-32. Bureau of Economic Analysis.

Mataloni, L., Avera, J., & Mayerhauser, N. (2014). *National income and product accounts gross domestic product: Second quarter 2014 (advance estimate) annual revision: 1999 through first quarter 2014.* No. BEA 14-31. Bureau of Economic Analysis.

Mataloni, L., & Hodge, A. (2010). *National income and product accounts gross domestic product, 2nd quarter 2010 (second estimate); corporate profits, 2nd quarter 2010 (preliminary estimate).* No. BEA 10-41. Bureau of Economic Analysis.

Mataloni, L., & Pinard, K. (2016). *Gross domestic product: Second quarter 2016 (second estimate); corporate profits: Second quarter 2016 (preliminary estimate).* No. BEA 16-44. Bureau of Economic Analysis.

McDonald, M. (2018). *United states elections project: Voter turnout.* Retrieved March 21, 2018, from http://www.electproject.org/home/voter-turnout/voter-turnout-data.

Murray, G. R., Riley, C., & Scime, A. (2009). Pre-election polling: Identifying likely voters using iterative expert data mining. *Public Opinion Quarterly, 73*(1), 159-171.

Newport, F. (2000). *How do you define "likely voters"?* Retrieved September 11, 2017, 2017, from http://www.gallup.com/poll/4636/how-define-likely-voters.aspx.

Paul Freedman, a., & Ken Goldstein, a. (1996). Building a probable electorate from preelection polls: A two-stage approach. *The Public Opinion Quarterly,* (4), 574.

Pew Research Center. (2016). *Can likely voter models be improved?*

Rentsch, A., & Schaffner, B. F. (Manuscript submitted for publication.). Weight for it: The limited role of weighting decisions on predicting polling errors in 2016.

Rogers, T., & Aida, M. (2014). Vote self-prediction hardly predicts who will vote, and is (misleadingly) unbiased. *American Politics Research, 42*(3), 503-528.

Sides, J. (2014). Four suggestions for making election forecasts better, and better known. *PS: Political Science and Politics, 47*(2), 339-341.

Silver, N. (2017, September 21). The media has a probability problem: The media’s demand for certainty — and its lack of statistical rigor — is a bad match for our complex world. *Fivethirtyeight.Com.*

Verba, S., & Nie, N. H. (1972). *Participation in America: Political Democracy and Social Equality.* Harper & Row.

Visser, P. S., Krosnick, J. A., Marquette, J., & Curtin, M. (2000). Improving election forecasting: Allocation of undecided respondents, identification of likely voters, and response order effects. In P. Lavrakas, & M. Traugott (Eds.), *Election polls, the news media, and democracy* (pp. 224-260). New York, NY: Chatham House.

Westwood, S., Messing, S., & Lelkes, Y. (2018). Projecting confidence: How the probabilistic horse race confuses and demobilizes the public. *Social Science Research Network.* April 26, 2018.

Wlezien, C., & Erikson, R. S. (2004). The fundamentals, the polls, and the presidential vote. *PS: Political Science & Politics, 37*(4), 747-751.

Wolfinger, R. E., & Rosenstone, S. J. (1980). *Who votes?* Yale University Press.

12. Appendix

A1. CCES question wordings and response options

Variable	Question wording	Response options
Vote intent	Do you intend to vote in [YEAR] [NAME OF ELECTION]?	• Yes, definitely • Probably • I already voted (early or absentee) • No • Undecided • I plan to vote before [ELECTION DATE] (2012 and 2014)
Vote history	Did you vote in the [PREVIOUS ELECTION YEAR] General Election? (2008: Did you vote in the Presidential primary or attend a caucus between January and June of this year?)	• No • I usually vote, but did not in [YEAR] (2010, 2012, 2014, 2016) • I am not sure (2010, 2012, 2014, 2016) • Yes (2008) • Yes. I definitely voted. (2010, 2012, 2014, 2016)
Political interest	Some people seem to follow what is going on in government or public affairs most of the time, whether there's an election or not. Others aren't that interested. Would you say you follow what is going on in government and public affairs...	• Most of the time • Some of the time • Only now and then • Hardly at all • Don't know
Voter registration status	Are you registered to vote?	• Yes • No • I don't know
Gender	Are you male or female?	• Male • Female
Age	In what year were you born?	--
Race	What racial or ethnic group best describes you?	• White • Black • Hispanic • Asian

		• Native American • Mixed • Other • Middle Eastern
Education	What is the highest level of education you have completed?	• No HS • High school graduate • Some college • 2-year • 4-year • Post-grad
Income	Thinking back over the last year, what was your family's annual income?	• Less than $10,000 • $10,000-$14,999 • $15,000-$19,999 • $20,000-$24,999 • $25,000-$29,999 • $30,000-$39,999 • $40,000-$49,999 • $50,000-$59,999 • $60,000-$69,999 • $70,000-$79,999 • $80,000-$99,999 • $100,000-$199,999 • $120,000-$149,999 • $150,000 or more • Prefer not to say • $150,000-$199,000 (2012 and 2014) • $200,000-$249,000 (2012 and 2014) • $250,000-$349,000 (2012 and 2014) • $350,000-$499,000 (2012 and 2014) • $250,000 or more (2012 and 2014) • $500,000 or more (2012 and 2014)
Partisanship	Generally speaking, do you think of yourself as a ...? [IF INDEPENDENT] Do you think of yourself as closer to the Democratic or Republican Party?	• Strong Democrat • Not very strong Democrat • Strong Republican • Not very strong Republican • Lean Democrat • Lean Republican

		• Independent • Not sure
Religiosity	Aside from weddings and funerals, how often do you attend religious services?	• More than once a week • Once a week • Once or twice a month • A few times a year • Seldom • Never • Don't know
Marital status	What is your marital status?	• Married • Separated • Divorced • Widowed • Single • Domestic partnership
Residential mobility	How long have you lived at your present address?	• Less than 1 month • 2 to 6 months • 7 to 11 months • 1 to 2 years • 3 to 4 years • 5 or more years

A2. Pew Perry-Gallup index question wordings and point assignment

Question wording	**Response options (point assignment)**
How much thought have you given to the upcoming November election?	• Quite a lot (+1) • Some (+1) • Only a little • None
Have you ever voted in your precinct or election district?	• Yes (+1) • No
Would you say you follow what's going on in government or public affairs?	• Most of the time (+1) • Some of the time (+1) • Only now and then • Hardly at all
How often would you say you vote?	• Always (+1) • Nearly always (+1) • Part of the time • Seldom

How likely are you to vote in the general election this November?	• Definitely will vote (+1) • Probably will vote (+1) • Probably will not vote • Definitely will not vote
In the 2012 presidential election between Barack Obama and Mitt Romney, did things come up that kept you from voting, or did you happen to vote?	• Yes, voted (+1) • No
Please rate your chance of voting in November on a scale of 10 to 1.	• 0 to 8 • 9 (+1) • 10 (+1)