FOR RELEASE JANUARY 7, 2016

Can Likely Voter Models Be Improved?

Evidence from the 2014 U.S. House elections

FOR MEDIA OR OTHER INQUIRIES:

Scott Keeter, Senior Survey Advisor
Ruth Igielnik, Research Associate
Rachel Weisel, Communications Associate

202.419.4372

www.pewresearch.org

RECOMMENDED CITATION: Pew Research Center, January, 2016, "Can Likely Voter Models Be Improved?"

About Pew Research Center

Pew Research Center is a nonpartisan fact tank that informs the public about the issues, attitudes and trends shaping America and the world. It does not take policy positions. The Center conducts public opinion polling, demographic research, content analysis and other data-driven social science research. It studies U.S. politics and policy; journalism and media; internet, science and technology; religion and public life; Hispanic trends; global attitudes and trends; and U.S. social and demographic trends. All of the center's reports are available at www.pewresearch.org. Pew Research Center is a subsidiary of The Pew Charitable Trusts, its primary funder.

Can Likely Voter Models Be Improved?

Evidence from the 2014 U.S. House elections

In recent years, polling has missed the mark in several high-profile elections, drawing particular attention to the difficulties inherent in using surveys to predict election outcomes. These failures typically result from one or more of three causes: biased samples that include an incorrect proportion of each candidate's supporters; change in voter preferences between the time of the poll and the election; or incorrect forecasts about who will vote. While not a new concern, the third of these – the difficulty of identifying likely voters – may be the most serious, and that is the focus of this study. Election polls face a unique problem in survey research: They are asked to produce a model of a population that does not yet exist at the time the poll is conducted, the future electorate.

It is well understood that many people who are eligible to vote and who tell pollsters they intend to cast a ballot will not actually do so. Similarly, some people who express little interest in the election or uncertainty about voting will nevertheless turn out. This is not a source of random error, because people who vote regularly are demographically and politically different from those who vote less often. In U.S. elections, experienced pollsters know that supporters of Republican candidates tend to be more likely to vote than supporters of Democratic candidates, especially in off-year elections. Consequently, identifying who is likely to vote is fundamental to making accurate forecasts from pre-election polls and correctly characterizing the views of the electorate.

This study examines various methods of determining who is a likely voter. It then compares the relative effectiveness of each approach in describing the electorate and measuring the division of the vote between parties in the 2014 U.S. House of Representatives elections. Pollsters would like to have a crystal ball that would allow them to see who will ultimately turn out to vote. While this study has no crystal ball, it has the next best thing: a survey of people interviewed before and after the 2014 congressional elections that is enhanced with verified turnout data from a national voter file (a database of adults and their publicly available voter turnout records from all states).

In particular, this study makes it possible to assess at least some of the benefits of sampling from lists of registered voters, the method favored by many campaign pollsters. Public pollsters, such as Pew Research Center and the major news organizations that conduct election polls, typically have used random digit dial (RDD) samples to reach a random sampling of all Americans, then narrowing down to prospective voters by asking people a series of questions that gauge interest in

the election, past voting behavior and intention to vote.[1] Campaign pollsters tend to use samples from databases of registered voters and incorporate past vote history from those databases into their forecasting models, ensuring that they know whether the respondent has voted in the past.[2] The sample employed in this study was originally obtained from an RDD survey and later matched to a voter file so that both the survey questions and the past vote history could be used in the analysis.

All of the methods examined here result in more-accurate forecasts than using either all those respondents who say they are registered to vote, or else all those who say they intend to vote, both of which include far too many people who ultimately will not cast a ballot. But some approaches performed better than others. Nearly all of the methods produced more-accurate forecasts when voter file records of previous voting were incorporated into the models.

How the study was conducted

The analysis is based on pre- and post-election interviews with 2,424 U.S. adults from Pew Research Center's nationally representative American Trends Panel who reported that they are registered to vote and were able to be matched to a national voter file. Panelists were interviewed from Sept. 9 to Oct. 3, 2014, about the upcoming congressional election. The survey included a range of standard questions about intention to vote, interest in the campaign, past voting experience and party preference in the election that Pew Research Center and others use to model the likely electorate. Panelists were re-interviewed from Nov. 17 to Dec. 15, 2014, and asked whether and for whom they voted in the election for the U.S. House of Representatives.

The names and addresses of most panelists were gathered as part of the core American Trends Panel methodology and used to match respondents from the survey sample to their corresponding record in a national voter file. To preserve the privacy of the panelists, the names and addresses of the panelists are securely stored and kept separate from the survey data and voter file information.

The voter file, gathered from publicly available individual voter lists from each state, contains information on nearly every voter's turnout history along with a variety of demographic information (the voter file does not indicate the candidates for whom a person voted, only whether he or she turned out in that election). Matching this voter file to our survey data allows us to

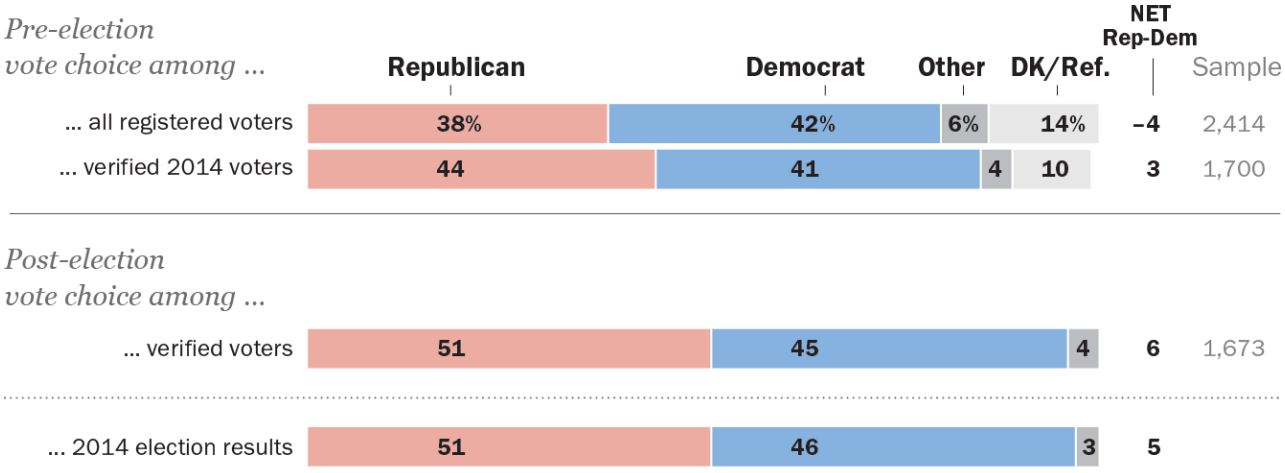
[1] An exception to this is polling in primaries. Some public polling organizations use voter files as sampling frames in primaries because they are far more efficient for reaching likely primary voters, who typically constitute a very small share of all voters.
[2] The exact methods employed by most public and campaign pollsters are proprietary and so could not be reproduced precisely here. Many campaign pollsters who sample from voter files also construct their samples to match what they expect to be the demographics and political characteristics of the likely electorate, rather than interviewing a broader group of voters and narrowing the sample to match the likely electorate.

incorporate past turnout history and to validate whether panel respondents were recorded as having cast a ballot in the 2014 contest.

1. Polls and votes: The 2014 elections by the numbers

Our equivalent of a crystal ball – the voter file, combined with a post-election survey interview – provides us with a validated record of turnout for our survey respondents. Our post-election survey provides us with the respondents' report of *how* they voted. This allows us to see how a Democratic advantage among registered voters in a survey conducted the September before the elections turned into a Republican win among those verified to have turned out to vote.

House vote choice among registered, self-reported and verified voters



Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter file. House vote choice based on the September wave conducted Sept. 9-Oct. 3, 2014.

PEW RESEARCH CENTER

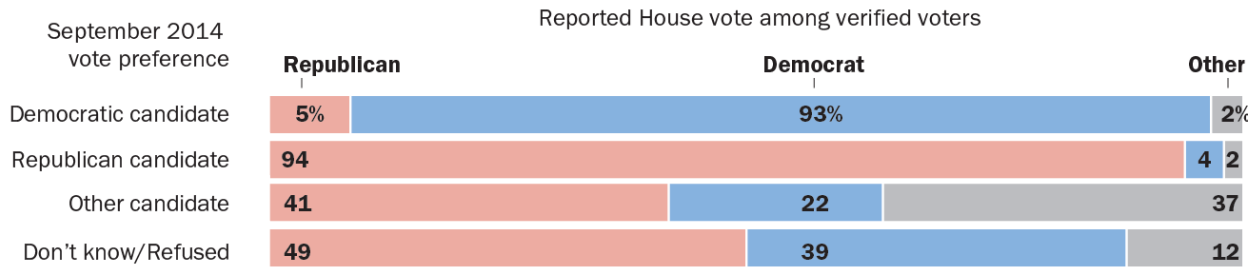
Of the three main reasons why election polls sometimes do not match the election results, we can rule out the first one, which is bias in the survey's sample. Among verified voters interviewed in the post-election survey, 51% reported voting Republican, 45% Democratic, almost exactly matching the outcome of the elections. This shows that the sample is not biased with respect to preferences in the elections for U.S. House.

The real answer for why the outcome differed from the pre-election estimate: Some people changed their minds, while others did not show up to vote. In 2014, at least, Republicans benefited from both of these factors: People who had thought of voting for a third-party candidate or were

undecided were more likely to shift toward the GOP, and those who ended up not voting disproportionately favored Democratic candidates.

In the September pre-election survey, 42% of registered voters favored a Democratic House candidate, while 38% favored a Republican, a 4 percentage point Democratic advantage.[3] If we could have used the perfect knowledge of hindsight, however, and only selected those who would eventually be verified as having actually turned out to vote, that same September survey would have found that Republican candidates held a 3-point lead at the time (44% vs. 41%). Given that the final GOP advantage among all tallied votes for the House of Representatives was nearly 6 points (51.4% vs. 45.7%),[4] the data suggest that correctly predicting who would turn out to vote would have produced a more accurate forecast, even when relying on candidate choices among voters more than one month prior to the elections.

Change in vote choice among panelists, pre- to post-election



Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter file.

PEW RESEARCH CENTER

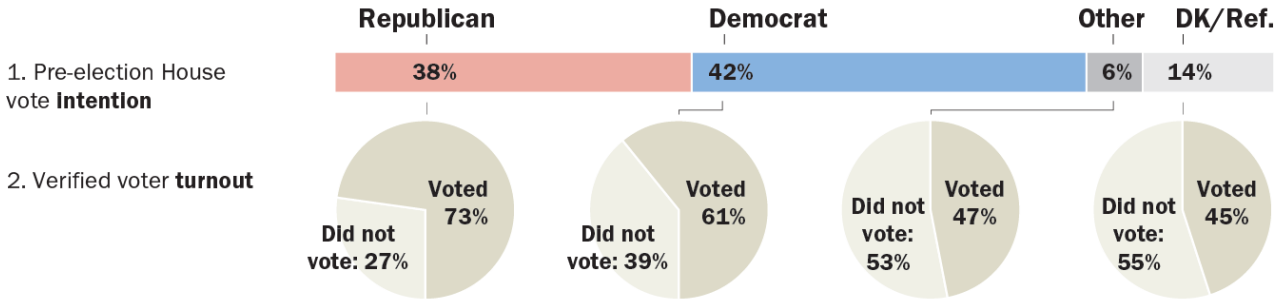
Since we interviewed these voters again after the elections, we can also tell whether they changed their minds in ways that affected the overall outcome. The vast majority of those supporting either a Republican or a Democrat in the pre-election poll remained loyal to their candidate (94% among Republicans, 93% among Democrats). But among the small number of voters supporting a third-party candidate or who had no preference in the pre-election waves, Republicans picked up more support than did the Democrats.

[3] Because these numbers are based on only those respondents who participated in both waves of the panel and who were able to be matched to the voter file, they may differ slightly from previously published results, which were based on all registered voters in each wave.
[4] The nationwide total vote for the House of Representatives in 2014, as compiled by David Wasserman of the Cook Political Report, showed that 51.4% voted for the Republican House candidate in their district, 45.7% voted for the Democratic candidate and 2.9% voted for another candidate.

While changed minds contributed to some of the difference between the September poll result and the final outcome, this factor was less important than the turnout differential between Republicans and Democrats. Fully 73% of pre-election registered voters who supported a Republican candidate in the pre-election survey ultimately turned out to vote on Election Day, based on verified vote from the voter file. By comparison, only 61% of registered voters who supported a Democratic candidate were verified to have voted.

Vote intention and turnout

Republican candidates benefited from higher turnout by their supporters. Among pre-election respondents who supported Republican candidates, 73% turned out to vote; among those who supported Democratic candidates, 61% turned out to vote.



Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter file.

PEW RESEARCH CENTER

As noted earlier, simply accurately identifying who would actually turn out to vote – without accounting for shifts in voter choices – would have shown the Republicans to be leading by about 3 percentage points in the pre-election survey, an improvement over the tally among all registered voters at the time (Republicans trailing by 4 points). On top of this, the GOP picked up an additional 2 points on the margin as a result of changing preferences from pre- to post-election.

In the next section, we examine a variety of approaches to distinguishing likely voters from nonvoters.

2: Measuring the likelihood to vote

The survey literature has long shown that more respondents say they intend to vote than actually cast a ballot (e.g., Bernstein et al. 2001; Silver et al. 1986). In addition, some people say they do not expect to vote but actually do, perhaps because they are contacted by a campaign or a friend close to Election Day and are persuaded to turn out. These situations potentially introduce error into election forecasts because these stealth voters and nonvoters often differ in their partisan preferences. In general, Republicans are more likely than Democrats to turn out, though they may be about equally likely to say they intend to vote. As a consequence, pollsters do not rely solely upon a respondent's stated intention when classifying a person as likely to vote or not. Instead, most ask several questions that collectively can be used to estimate an individual's likelihood of voting. The questions measure intention to vote, past voting behavior, knowledge about the voting process and interest in the campaign.

This study examines different ways of using seven standard questions, and sometimes other information, to produce a model of the likely electorate. The questions were originally developed in the 1950s and '60s by election polling pioneer Paul Perry of Gallup and have been used – in various combinations and with some alterations – by Pew Research Center, Gallup and other organizations in their pre-election polling (Perry 1960, 1979). The questions tested here include the following (the categories that give a respondent a point in the Perry-Gallup index, discussed in the following section, are in **bold**):

- How much thought have you given to the coming November election? ***Quite a lot, some,*** *only a little, none*
- Have you ever voted in your precinct or election district? ***Yes,*** *no*
- Would you say you follow what's going on in government and public affairs ***most of the time, some of the time,*** *only now and then, hardly at all?*
- How often would you say you vote? ***Always, nearly always,*** *part of the time, seldom*
- How likely are you to vote in the general election this November? ***Definitely will vote, probably will vote,*** *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;*** *no*
- Please rate your chance of voting in November on a scale of 10 to 1. *0-8,* ***9, 10***

Some pollsters have employed other kinds of variables in their likely voter models, including demographic characteristics, partisanship and ideology. Below we evaluate models that use these types of measures as well.

Two additional kinds of measures tested here are taken from a national voter file. These include indicators for past votes (in 2012 and 2010) and a predicted turnout score that synthesizes past voting behavior and other factors to produce an estimated likelihood of voting. These measures are strongly associated with voter turnout. A detailed analysis of all of these individual measures and how closely each one is correlated with voter turnout and vote choice can be found in Appendix A to this report.

Two broad approaches are used to produce a prediction of voting with pre-election information such as the Perry-Gallup questions or self-reported past voting history (Burden 1997). *Deterministic* methods use the information to categorize each survey respondent as a likely voter or nonvoter, typically dividing voters and nonvoters using a threshold or "cutoff" that matches the predicted rate of voter turnout in the election. *Probabilistic* methods use the same information to compute the probability that each respondent will vote. Probabilities can be used to weight respondents by their likelihood of voting, or they can be used as a basis for ranking respondents for a cutoff approach. This analysis examines the effectiveness of both approaches.

The Perry-Gallup likely voter index

Following the original method developed by Paul Perry, Pew Research Center combines the individual survey items to create a scale that is then used to classify respondents as likely voters or nonvoters. For each of the seven questions, a respondent is given 1 point for selecting certain response categories. For example, a response of "yes" to the question "Have you ever voted in your precinct or election district?" gets 1 point on the scale. Younger respondents are given additional points to account for their inability to vote in the past (respondents who are ages 20-21 get 1 additional point and respondents who are ages 18-19 get 2 additional points).[5] Additionally, those who say they "definitely will not" be voting, or who are not registered to vote, are automatically coded as a zero on the scale. As tested here, the procedure results in an index with values ranging from 0 to 7, with the highest values representing those with the greatest likelihood of voting.

The next step is to make an estimate of the percentage of the eligible adults likely to vote in the election. This is typically based on a review of past turnout levels in similar elections, adjusted for judgments about the apparent level of voter interest in the current campaign, the competitiveness of the races and degree of voter mobilization underway. The estimate is used to produce a "cutoff" on the likely voter scale, selecting the highest-scoring respondents based on the expected turnout in the coming election. For example, if we expected that 40% of the voting eligible population would vote (a typical turnout for a midterm election), then we would base our survey estimates on

[5] The application of bonus points is handled slightly differently in presidential vs. off-year elections.

the 40% of the eligible public receiving the highest index scores.[6] In reality, 36% of the eligible adult population turned out in 2014. The choice of a turnout threshold is a very important decision because the views of voters and nonvoters are often very different, as was the case in 2014. (See Appendix C for data on how the choice of a turnout target matters.)[7]

SIDEBAR BOX with title "What if the survey includes too many politically engaged people?": One complication with the application of a turnout estimate to the survey sample is the fact that election polls tend to overrepresent politically engaged individuals. It may be necessary to use a higher turnout threshold in making a cutoff to account for the fact that a higher percentage of survey respondents than of members of the general public may actually turn out to vote. Unfortunately, there is no agreed-upon method of making this adjustment, since the extent to which the survey overrepresents the politically engaged, or even changes the respondents' behavior (e.g., by increasing their interest in the election), may vary from study to study and is difficult to estimate.

The data used here include only those who are registered to vote; consequently, the appropriate turnout estimate within this sample should be considerably higher than among the general public. For many of the simulations presented in this report, we estimated that 60% of registered voters would turn out. Assuming that 70% of adults are registered to vote, this would equate to a prediction of 42% turnout of the general public.[8]

In these data, an expectation of 60% turnout meant that all respondents who scored a 7 on the scale (48% of the total) would be classified as likely voters, along with a weighted share of those who scored 6 (who were 15% of the total).

Deterministic (or cutoff) methods like this one leave out many actual voters. While those coded 6 and 7 on the scale are very likely to vote (63% and 83% of each group, respectively were validated to have voted), there also are many actual voters among those who scored below 6: About a fifth (22%) of all verified voters scored between 0 and 5. Of course, the goal of the model is not to classify every respondent but to produce an accurate aggregation of the vote. But if the distribution of those correctly classified does not match that of the actual electorate, the election forecast will be wrong.

---

[6] To reach a precise percentage of "likely voters" from the 7-item index, it is often necessary to take all respondents from the one or two highest-scoring categories and a weighted proportion of the next category. For example, if we estimated a turnout of 50%, and 40% of the sample scored a 7 on the index and 15% scored a 6, we would count all of the "7"s and weight the category "6" by 0.6667 (10/15) to estimate the likely voter pool.

[7] The voting-age population is slightly different from the voting-eligible population. The latter excludes noncitizens and citizens who have lost the right to vote. Many among the latter group are not living in residential households and thus are unlikely to be reachable by the typical election poll.

[8] Based on the average of Pew Research Center estimates from telephone surveys conducted in fall 2014.

Consistent with general patterns observed in previous elections of this type, respondents who scored a 7 on the scale favor Republican over Democratic candidates (by a margin of 50% to 44%). Majorities of those in categories 5 and 6 prefer Democratic candidates. As in most elections, the partisan distribution of the predicted vote depends heavily on where the line is drawn on the likely voter scale. Including more voters usually makes the overall sample more Democratic, especially in off-year elections. That is why judgments about where to apply the cutoff are critical to the accuracy of the method.

The Perry-Gallup index

Score on scale	Share of registered voters	Share of all verified voters	% who are verified voters in each group
7	48%	63%	83%
6	15	15	63
5	10	10	59
4	7	4	34
3	6	4	41
2	6	2	23
1	3	1	13
0	4	1	11
	100	100	

83% of those who scored a 7 actually voted

22% of verified voters scored between 0 and 5

Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter.

PEW RESEARCH CENTER

Probabilistic models

The same individual survey questions can also be used to create a statistical model that assigns a predicted probability of voting to each respondent, along with coefficients that measure how well each item correlates with turnout. These coefficients can then be used in other elections with surveys that ask the same questions to create a predicted probability of voting for each respondent, based on the assumption that expressions of interest, past behavior and intent all have the same impact regardless of the election. All response options for each item can be used in the model, or they can be coded as they are in the Perry-Gallup method. Regardless of the form of the inputs, the result is a distribution, with each respondent assigned a score on a scale corresponding to how likely he or she is to turn out to vote. If someone is classified as a 0.30, then that respondent is thought to have a 30% chance of voting.

One potential benefit to this method is that it can use more of the information contained in the survey (all of the response categories in each question, rather than just a selected one or two). This also gives respondents who may have a lower likelihood of voting – whether because of their age, lack of ongoing interest in the election or simply having missed a past election – a possibility of affecting the outcome, since we know that many who score lower on the scale actually do vote.

These respondents will be counted as long as they have a chance of voting that is greater than zero; they are simply given a lower weight in the analysis than others with a higher likelihood of voting.

One potential drawback of this method is that it applies a model developed in a previous election to a current election, based on the assumption that the relationships between turnout and the key predictors are the same across elections. In this study, our models are built using voter participation data from the 2014 elections, and the resulting weights are applied retroactively to produce survey estimates of the likely 2014 vote. As a result, we cannot test how well these models would perform in future elections. The likely voter model used by CBS News, which has employed a variation of this method for decades, suggests that such assumptions are reasonable. Rather, our goal is to explore the differences between probabilistic and deterministic approaches to modeling voter turnout, and learn how much these models are improved when we include information on prior voting behavior drawn from the voter file.

In our evaluations of probabilistic models, we also tested a "kitchen sink" model that includes the seven Perry-Gallup measures along with a range of demographic and political variables including age, education, income, race/ethnicity, party affiliation, ideological consistency, home ownership and length of tenure at current residence – all factors known to be correlated with voter turnout.

In testing probabilistic approaches, we explored two methods for creating predicted probabilities: logistic regression, a common modeling tool, and a machine-learning technique known as "random forest."

In addition to using the predicted probabilities as a weight, they can also be used with a cutoff. As with the Perry-Gallup scale, the cutoff method would count the top-scoring respondents as likely voters and ignore the others. For example, assuming that 60% of registered voters are going to turn out, the models would include only the top 60% of respondents as ranked by their predicted probabilities of voting.

Logistic regression

To build a model comparable to the Perry-Gallup 7-item scale, the same seven questions on voter engagement, past voting behavior, voter intent and knowledge about where to vote were used. (The "kitchen sink" model used these items along with demographic and political variables.) The questions were entered into the model as predictors without combining or collapsing categories. Variables were rescaled to vary between 0 and 1, with "don't know" responses coded as zero.

A logistic regression was performed using verified vote from the voter file as the dependent variable. The regression produces a predicted probability of voting for each respondent and coefficients for each measure. The probabilities are then used in various ways as described below to produce a model of the electorate for forecasting. In subsequent elections, the coefficients derived from these models can be used with the answers from respondents in contemporary surveys to produce a probability of voting for each person. As with the Perry-Gallup approach, this method assumes that the measures used in the study are equally relevant for distinguishing voters from nonvoters in a variety of elections.

Decision trees and random forests

Another probabilistic approach involves the use of "decision trees" to identify the best configuration of variables to predict a particular outcome – in this case, voting and nonvoting. The typical decision tree analysis identifies various ways of splitting a dataset into separate paths or branches, based on options for each variable. The decision tree approach can be improved using a machine-learning technique known as "random forests." Random forests employ large numbers of trees fit to random subsamples of the data in order to provide more precise predictions than would be obtained by fitting a single tree to all of the data. Unlike classical methods for estimating probabilities such as logistic regression, random forests perform well with large numbers of predictor variables and in the presence of complex interactions. We applied the random forest method to the computation of vote probabilities, starting with the same variables employed in the other methods described earlier.

When a single decision tree is fit to a dataset, the algorithm starts by searching for the value among the predictor variables that can be used to split the dataset into two groups that are most homogenous with respect to the outcome variable, in this case whether or not someone voted in the 2014 elections. These subgroups are called nodes, and the decision tree algorithm proceeds to split each node into progressively more and more homogenous groups until a stopping criterion is reached. One thing that makes the random forest technique unique is that prior to splitting each node, the algorithm selects a random subset of the predictor variables to use as candidates for splitting the data. This has the effect of reducing the correlation between individual trees, which further reduces the variance of the predictions.

When employing statistical models for prediction, it is important to address the possibility that the models are overfitting the data – finding patterns in data that reflect random noise rather than meaningful signal – which reduces their accuracy when applied to other datasets. This is less of a concern for logistic regression, which is unlikely to overfit when the sample size is large relative to the number of independent variables (as is the case here). But it *is* a concern for powerful

machine-learning methods such as random forests that actively seek out patterns in data. One advantage of random forests in this regard is the fact that each tree is built using a different random subsample of the data. In our analysis, the predicted probabilities for a case are based only on those trees that were built using subsamples where that case was excluded. The result is that any overfitting that occurs in the tree-building process does not carry over into the scores that are applied to each case.

One final regression-based method tested here is to employ a voter turnout probability created by the voter file vendor as a predictor or a weight. The TargetSmart voter file includes a 2014 turnout likelihood score developed by Clarity Campaign Labs. This score ranges from 0 to 1 and can be interpreted as a probability of voting in the 2014 general election.

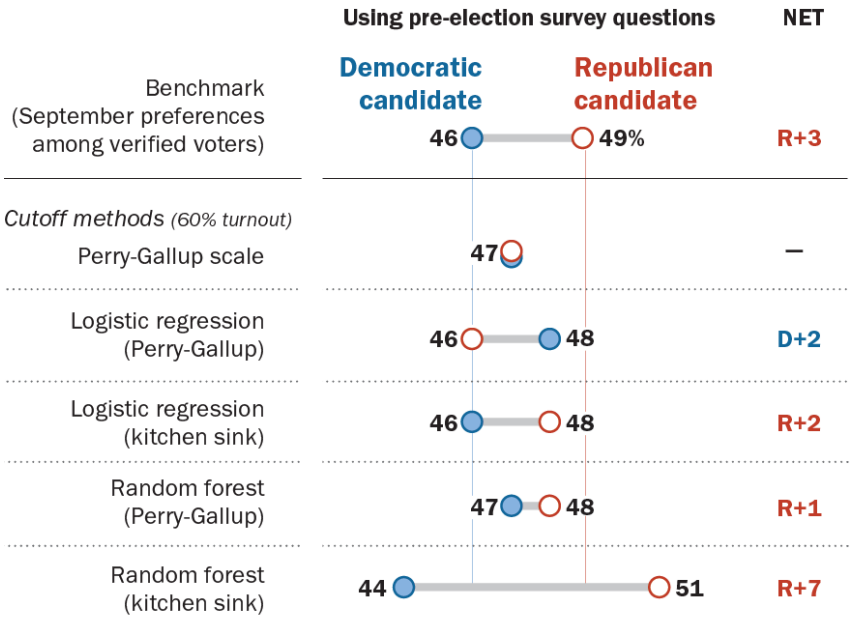
The statistical analysis reported in the next section uses the verified vote as the measure of turnout. Among registered voters in the sample, 63% have a voter file record indicating that they voted in 2014. Self-reported voting was more common; 75% of registered voters said they turned out. Appendix B discusses the pros and cons of using verified vote vs. self-reported vote.

3. Comparing the results of different likely voter models

All told, we tested 16 different variations on four types of likely voter methods, producing estimates that range from a 2-point Democratic lead to a 7-point Republican advantage in the generic U.S. House vote. The benchmark for comparison is a 3-point Republican lead among verified voters (49% Republican, 46% Democratic) when they were interviewed prior to the elections. While there is no objective way to know where the race stood at the time the September poll was conducted, this benchmark is our comparison of choice. With hindsight and a voter file, we are able to know what voter preferences a perfect likely voter measure – one that included 100% of eventual voters and zero nonvoters – would have produced with this survey.

Estimates using cutoff methods

September pre-election voter preferences



Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter.

PEW RESEARCH CENTER

Cutoff methods

Using a 60% registered voter turnout cutoff (projecting a 42% overall turnout), the Perry-Gallup index yielded a tied race, at 47% for each party's candidate. A more accurate prediction can be obtained if a lower turnout forecast is used (see Appendix C), but given that the actual turnout in the sample was 63%, a tighter turnout screen would be inappropriate.

A logistic regression to produce probabilities of turnout, using the standard seven likely voter questions with a 60% cutoff, produced a 2-point Democratic edge (48% vs. 46% Republican).

Adding demographic and political variables from the pre-election survey (the "kitchen sink") improved the predictions, yielding a 2-point Republican advantage.

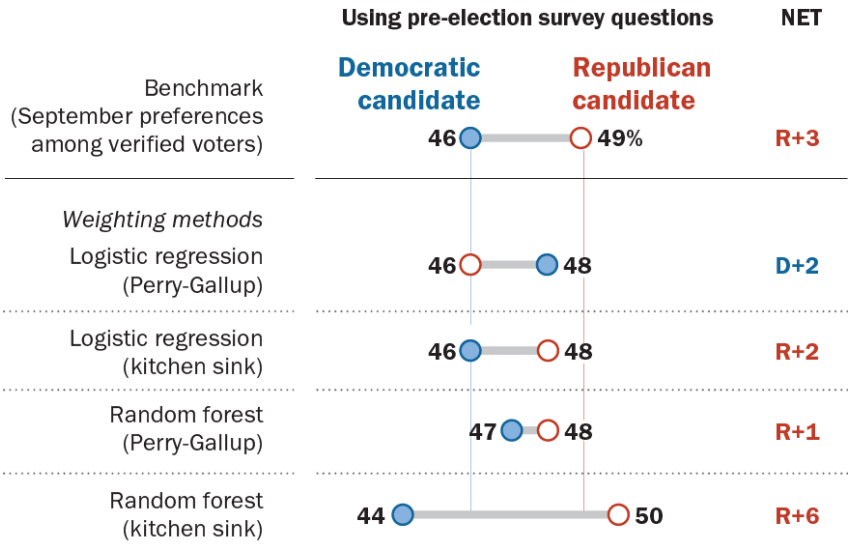
Probabilities produced with a random forest model (using the Perry-Gallup items and the same 60% cutoff) produced a 1-point GOP advantage (48% to 47%). Adding the demographic and political variables drove the forecast well past the benchmark, to a 7-point GOP lead.

Weighting methods

Using predicted vote probabilities as weights produced results similar to those from the cutoff method. As noted earlier, using the probabilities as a weight as opposed to using a cutoff means including all respondents, even those who may have a lower likelihood of voting; they are simply given a lower weight in the analysis than others with a higher likelihood of voting.

Estimates using weighting methods

September pre-election voter preferences



Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter.

PEW RESEARCH CENTER

Weights computed from logistic regression with the basic Perry-Gallup variables resulted in a forecast of a 2-point Democratic advantage (48% to 46%). Adding the kitchen sink of demographic and political variables shifted the estimate to a 2-point GOP edge.

Computing the probabilities with the random forest method produces a 1-point GOP advantage (48% to 47%), while adding the kitchen sink produced a 6-point GOP advantage, 3 points larger than the benchmark.

Adding voter file history

Adding voter file records of verified past turnout history for the 2010 midterm elections and the 2012 presidential elections widens or creates a Republican advantage for nearly every prediction, in most cases improving estimates of the 2014 vote compared with the benchmark.

In particular, adding evidence of past turnout to the logistic regression using the Perry-Gallup scale and the 60% cutoff produced a much more accurate result, turning a 2-point Democratic advantage into a 2-point Republican advantage.

Adding the vote history variables to the traditional Perry-Gallup scale with a 60% cutoff made less of a difference. With vote history, this approach produced a 1-point GOP edge (48% GOP, 47% Democratic); without those variables, the model produced a tie (47% to 47%).

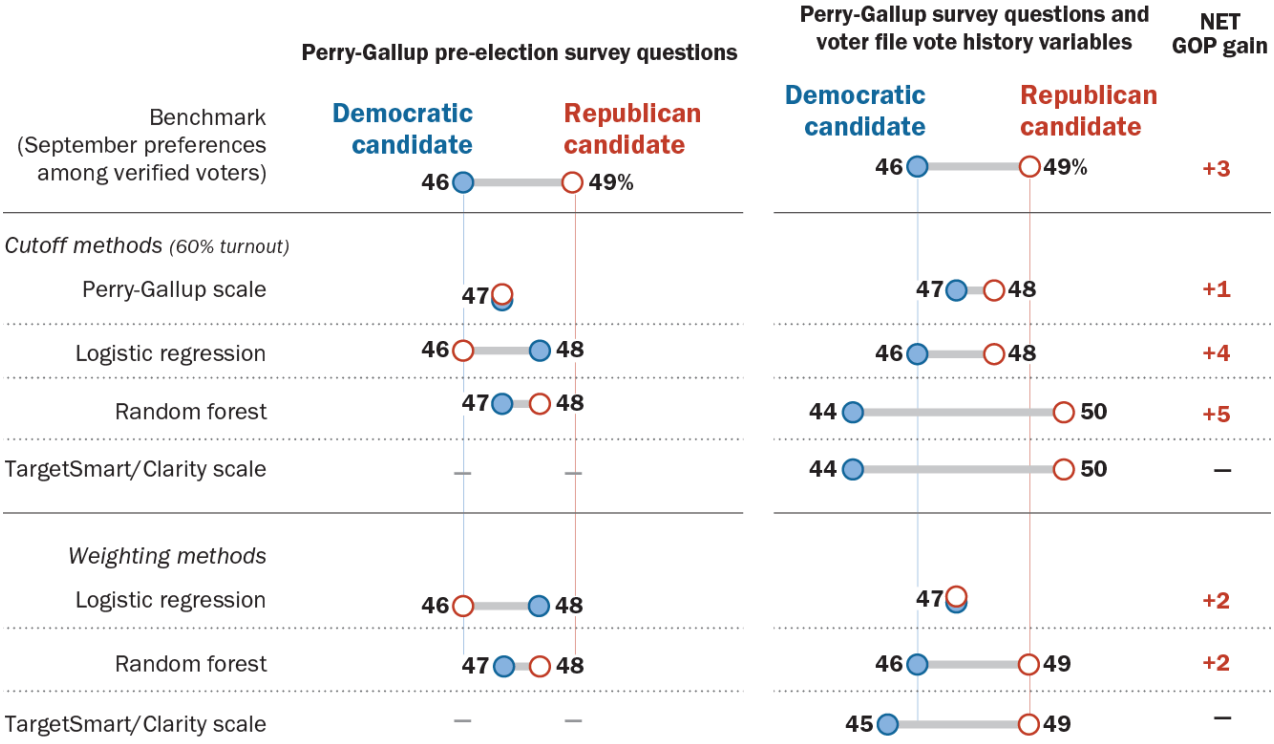
Adding vote history from the voter file to the random forest model estimates produced an estimate that was 2 points more Republican than the benchmark (50% Republican, 44% Democrat).

When vote history variables are included in models and employed as a weight, similar improvements are observed. Including the vote history in the random forest model produces a forecast that matches the benchmark (49% Republican, 46% Democratic). With the logistic regression, the vote margin moves from a 2-point Democratic lead (without the vote history variables) to a tie (47% for both). The TargetSmart/Clarity turnout score, which includes past vote history from the voter file, produced a 4-point GOP advantage when used as a weight.

It's important to keep in mind that the probabilistic methods tested here have an advantage over the original Perry-Gallup method, in that they have been computed using verified turnout in the current election as the dependent variable. While they produce more-accurate results in this test than the traditional Perry-Gallup method, their broader applicability depends on the assumption that the survey measures will be related to voter turnout in other elections in the same way that they are in 2014. By contrast, the Perry-Gallup approach has proven to be accurate in a wide range of elections for several decades.

Voter file vote history improves most estimates

September pre-election voter preferences



Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter.

PEW RESEARCH CENTER

Comparing the demographic profiles of likely electorates

When election pollsters apply a likely voter screen to their data, they look at more than its impact on the candidate preference question. They are also interested in what that model suggests about the makeup of the likely electorate in terms of key demographic groups. The different methods examined here produce likely electorates that are similar in terms of gender, race and age. In addition, they were not notably different on key characteristics – such as the share that is black or Hispanic, or ages 65 and older – from the profiles of the 2014 electorate produced by the 2014 national exit poll and the Voting and Registration Supplement conducted by the U.S. Census Bureau. The exit poll finds a somewhat better-educated and more affluent electorate than our survey. In particular, two-in-ten voters (20%) validated through the national voter file report family incomes of $100,000 or more annually, compared with 30% of voters in the national exit poll and 28% in the Census survey. This difference is potentially consequential because more affluent voters were considerably more Republican in their vote preferences than were the less affluent.

Demographic profile of voters and likely voters

	General public (adults)	2014 voters (exit poll)	2014 voters (CPS)	Verified voters	Perry-Gallup likely voters (60% cutoff)	Weighted based on logistic regression	Weighted based on random forest
	%	%	%	%	%		
Men	48	49	47	48	49	48	49
Women	52	51	53	52	51	52	51
White	65	75	76	75	76	74	75
Black	12	12	12	11	10	11	10
Hispanic	15	8	7	9	8	9	8
18-24	12	7	5	5	5	6	5
25-29	9	6	5	6	4	6	5
30-39	17	13	13	13	13	14	13
40-49	17	19	16	17	17	17	17
50-64	26	33	33	33	34	32	33
65+	19	22	28	26	27	25	27
Postgrad	11	20	16	13	13	12	13
College grad	19	31	25	24	24	23	23
Some college	29	29	30	32	33	34	34
High school or less	41	20	29	30	30	30	30
Family income							
$100,000+	22	30	28	20	19	19	19
$50,000-$99,999	30	34	34	33	32	32	32
$30,000-$49,999	20	20	18	19	19	19	19
Under $30,000	28	16	19	23	24	26	24

Source: National Election Pool 2014 national exit poll (for exit poll demographics).Voting and Registration Supplement, Current Population Survey, November 2014 (for CPS voter demographics). 2014 American Trends Panel September and November waves (for all others). Panel data based on registered voters who participated in both waves and were matched to a national voter file.

PEW RESEARCH CENTER

4. Conclusion

The analysis presented here suggests that modeling the electorate is likely to continue to vex pollsters, especially if no official record of past voting is available as an input to the models. As if to affirm this somewhat pessimistic conclusion, polls have failed to accurately predict winning candidates in several recent elections, including the 2015 race for governor in Kentucky, several 2014 U.S. races for Senate and governor, the 2015 British general election, the 2015 Scottish referendum on independence and the 2015 referendum in Greece on acceptance of the European Union's terms for a bailout. In the 2012 U.S. presidential election, many surveys at both the state and national levels underestimated the share of the vote that Barack Obama would receive. Errors in modeling the likely electorate are suspected of contributing to many of these polling failures.

So, can likely voter models be improved? For the 2014 elections, this analysis found that the Perry-Gallup method improved the U.S. House forecast relative to relying on the preferences of all registered voters, or even the subset who simply said they intended to vote in the election. But it did not perform as well as other approaches that incorporated more variables or more-complex models. A new modeling approach that uses decision trees and machine learning with the same set of questions improved on the estimates, but may be difficult for most polling organizations to implement and describe to their audiences. Moreover, it remains to be seen how well the regression methods evaluated here (including those using machine learning) will perform when they are applied to future elections.

Consistent with previously published research (e.g., Rogers and Aida 2014), adding voter file records of past vote produced the greatest improvement in the forecasts. But this information is often difficult to incorporate with random digit dial phone surveys since it requires gathering respondent names and addresses to facilitate an accurate match with the voter file; many RDD survey respondents are unwilling to provide this type of information during a telephone interview.

One solution is to use voter files as a sampling frame. This is becoming more common as the quality of state voter files and the commercial databases built upon them has improved. These commercial files often include telephone numbers and additional political, demographic and lifestyle data about households. But they may have significant biases, with highly mobile and lower-income individuals underrepresented (Jackman and Spahn 2015a, 2015b).

The ultimate goal of likely voter methods is to create an accurate model of the electorate, not necessarily to identify whether each individual in a survey will or will not vote. All of the methods examined here yield a predicted electorate that closely matches the actual 2014 electorate with respect to gender, age and race – three demographic variables that are correlated with vote choice.

Some pollsters adjust their models by making assumptions about the turnout of groups within the population or by attempting to match the characteristics of previous electorates. But if those assumptions are incorrect, serious forecasting errors can occur, as some GOP pollsters discovered when turnout among African Americans in 2012 exceeded their predictions.

Elections remain unique among the subjects that polls engage, in part because they provide a definitive outcome with which to judge the accuracy of the polls. Considering the precision that is required and the challenges inherent in election polling, it is perhaps notable that the craft has been as successful as it has. National polls in U.S. presidential elections in the past several cycles have been generally accurate in forecasting the partisan division of the vote, and state-level polls in 2012 were accurate enough to allow polling aggregators to forecast the outcome of the vote in all 50 states. But off-year elections in the U.S., especially 2014, were less kind to pollsters, and more recent national and international elections have raised further questions about whether polling is still able to accurately identify the electorate and its intentions.

The models examined here will need to be tested in future elections. Having panel data with information to validate turnout provides a strong basis for inference, but this is, in effect, a case study: one analysis in one election. This research focused on a particular midterm election, one that happened to have an unusually low turnout. Applying these models to other elections will reveal how well they can be generalized.

Acknowledgements

This report is a collaborative effort based on the input and analysis of the following individuals:

Primary Researchers
Scott Keeter, *Senior Survey Advisor*
Ruth Igielnik, *Research Associate*
Andrew Mercer, *Research Methodologist*
Jocelyn Kiley, *Associate Director, Research*

Collaborating Researchers
Claudia Deane, *Vice President, Research*
Michael Dimock, *President*
Ken Goldstein, *University of San Francisco*
Courtney Kennedy, *Director of Survey Research*
Amanda Lee, *Intern, Methodology*
Nick Hatley, *Research Assistant, Methodology*
Meredith Dost, *Research Assistant, Politics*
Hannah Fingerhut, *Research Assistant, Politics*
Elisa Shearer, *Research Assistant, Journalism*
Kyley McGeeney, *Research Methodologist*

Editorial and Graphic Design
Bill Webster, *Information Graphics Designer*
David Kent, *Copy Editor*

Communications and Web Publishing
Rachel Weisel, *Communications Associate*
Shannon Greenwood, *Assistant Digital Producer*
Travis Mitchell, *Digital Producer*

Methodology

The American Trends Panel surveys (ATP)

The American Trends Panel (ATP), created by Pew Research Center, is a nationally representative panel of randomly selected U.S. adults living in households. Respondents who self-identify as internet users (representing 89% of U.S. adults) participate in the panel via monthly self-administered Web surveys, and those who do not use the internet participate via telephone or mail. The panel is being managed by Abt SRBI.

Data in this report are drawn from two waves of the panel, September and November. The September wave was conducted Sept. 9-Oct. 3, 2014, among 3,154 respondents (2,811 by Web and 343 by mail). The November wave was conducted Nov. 17-Dec. 15, 2014, among 3,212 respondents (2,856 by Web and 356 by mail). For the purpose of this report, respondents are included only if they responded to both the September and November waves of the panel, told us they were registered to vote, and were able to be matched to the national voter file (a total of 2,424 respondents). The margin of sampling error for the full sample of 2,424 respondents is plus or minus 2.3 percentage points.

All current members of the American Trends Panel were originally recruited from the 2014 Survey of Political Polarization, a large (n=10,013) national landline and cellphone random digit dial (RDD) survey conducted Jan. 23 to March 16, 2014, in English and Spanish. At the end of that survey, respondents were invited to join the panel. The invitation was extended to all respondents who use the internet (from any location) and a random subsample of respondents who do not use the internet.[9]

Of the 10,013 adults interviewed, 9,809 were invited to take part in the panel. A total of 5,338 agreed to participate and provided either a mailing address or an email address to which a welcome packet, a monetary incentive and future survey invitations could be sent. Panelists also receive a small monetary incentive after participating in each wave of the survey.

The ATP data were weighted in a multi-step process that begins with a base weight incorporating the respondents' original survey selection probability and the fact that some panelists were subsampled for invitation to the panel. Next, an adjustment was made for the fact that the propensity to join the panel varied across different groups in the sample, as well as to correct for differences between the adults who completed the September and November waves and the adults

[9] When data collection for the 2014 Political Polarization and Typology Survey began, non-internet users were subsampled at a rate of 25%, but a decision was made shortly thereafter to invite all non-internet users to join. In total, 83% of non-internet users were invited to join the panel.

who did not complete both waves (either because they declined to join the panel, joined the panel but dropped out, or are still active in the panel but did not complete both waves). The final step in the weighting uses an iterative technique that matches gender, age, education, race, Hispanic origin, telephone service, population density and region to parameters from the U.S. Census Bureau's 2012 American Community Survey. It also adjusts for party affiliation using an average of the three most recent Pew Research Center general public telephone surveys, and for internet use using as a parameter a measure from the 2014 Survey of Political Polarization. Sampling errors and statistical tests of significance take into account the effect of weighting. The Hispanic sample in the ATP is predominantly native-born and English speaking. In addition to sampling error, one should bear in mind that question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of opinion polls.

Of the confirmed members of the panel, 69% responded to both the September and November waves. Taking into account the response rate for the 2014 Survey of Political Polarization (10.6%), the cumulative response rate for the September and November respondents is 3.1%.

Voter file matching

The names and addresses of most panelists were gathered as part of the core American Trends Panel methodology and used to match respondents from the survey sample to their corresponding record in a national voter file. The voter file, gathered by TargetSmart from publicly available individual voter lists from each state, contains information on most voters' turnout history and selected demographic information (note that the voter file does not indicate for which candidate a person voted, only whether they turned out in that election). To match panelists to the voter file, TargetSmart first looked for exact matches using name, address, and demographic characteristics. A second attempt was made with proximity matching, where a radius is drawn around the given address to test slight variations on the match. In total, 89% of respondents from the September and November waves of the panel were matched to the national voter file.

References

Ansolabehere, Stephen, and Eitan Hersh. 2012. "Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate." *Political Analysis* 20:437-59.

Bernstein, Robert, Anita Chadha, and Robert Montjoy. 2001. "Overreporting Voting: Why it Happens and Why it Matters. *Public Opinion Quarterly* 65:22-44.

Berent, M. K., 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: American National Election Studies. Retrieved from http://www.electionstudies.org/resources/papers/nes012554.pdf

Burden, Barry. 1997. "Deterministic and Probabilistic Voting Models." *American Journal of Political Science* 41:1150-69.

Colasanto, Diane and Jay A. Mattlin. 1987. "Evaluation of Gallup's Methodology for Predicting Likelihood of Voting." Paper prepared for presentation at the 1987 Joint Statistical meetings, San Francisco, Aug. 17, 1987.

Crespi, Irving. 1988. *Pre-election Polling: Sources of Accuracy and Error*. New York: Russell Sage Foundation.

Dimock, Michael, Scott Keeter, Mark Schulman and Carolyn Miller. 2001. "Screening for Likely Voters in Pre-election Surveys." Presented at the Annual Meeting of the American Association for Public Opinion Research, Montreal, May 17-20.

Freedman, Paul and Ken Goldstein. 1996. "Building a Probable Electorate from Pre-election Polls: A Two-Stage Approach." *Public Opinion Quarterly* 60:574-87.

Jackman, Simon and Bradley Spahn. 2015a. "Silenced and Ignored: How the Turn to Voter Registration Lists Excludes People and Opinions from Political Science and Political Representation." Unpublished paper. Accessed Dec. 22, 2015, at https://www.dropbox.com/s/qvqtz99i4bhdore/silenced.pdf?dl=0

Jackman, Simon and Bradley Spahn. 2015b. “Unlisted in America.” Unpublished paper. Accessed Dec. 22, 2015, at https://www.dropbox.com/s/bv5z1tyv9q422aw/Jackman%20Spahn%20-%20Unlisted%20in%20America.pdf?dl=0

Perry, Paul. 1960. “Election Survey Procedures of the Gallup Poll.” *Public Opinion Quarterly* 24:531-42.

Perry, Paul. 1979. “Certain Problems in Election Survey Methodology.” *Public Opinion Quarterly* 43:312-25.

Rogers, Todd and Masahiko Aida. 2013. “Vote Self-Prediction Hardly Predicts Who Will Vote, and Is (Misleadingly) Unbiased.” *American Politics Research,* 503-528.

Silver, Brian D., Barbara A. Anderson, and Paul R. Abramson. 1986. “Who Overreports Voting?” *American Political Science Review* 80:613-24.

Zukin, Cliff. 2004. “Sources of Variation in Published Election Polling: A Primer.” American Association for Public Opinion Research. Available at http://aapor.org/pdfs/varsource.pdf

Appendix A: The Perry-Gallup measures

The items used in the so-called Perry-Gallup scale – originally developed in the 1950s and '60s by election polling pioneer Paul Perry of Gallup and used in various combinations and with some alterations by the Pew Research Center, Gallup and other organizations in their pre-election polling (Perry 1960, 1979) – are widely employed by survey researchers in part or whole. This appendix presents data on how well each item discriminated between voters and nonvoters in 2014.

Measures of vote intention

The most direct way of predicting voter turnout is to simply ask whether a person intends to vote or not. And, in fact, this is the likely voter method that many pollsters use: Those respondents who say they will vote are included in the survey; those who say they won't vote are not. There are a variety of different ways of asking the "do you plan to vote" item. Since the vast majority of registered voters (90% or more) say they plan to vote, many researchers ask a follow-up question to gauge certainty of voting among those who plan to vote.

Measures of vote intention

	Share of total	% who are verified voters	*Vote intention (pre-election)* Republican	Democratic
	%	%	%	%
Likelihood of voting				
Definitely will vote	70	77	46	48
Probably will vote	20	36	36	54
Probably will not vote	7	13	33	57
Definitely will not vote	2	15	80	16
	100			
Chance of voting				
9-10 (higher likelihood)	75	75	46	48
7-8	11	34	37	50
5-6	5	27	26	65
3-4	4	16	47	37
1-2 (lower likelihood)	4	8	43	55
	100			

Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter file. Categories in bold contribute to the likely voter index.

PEW RESEARCH CENTER

In the American Trends Panel, we asked a slightly different question about the respondent's likelihood of voting: 70% of registered voters told us they "definitely" would be voting in the 2014 election. More than three-quarters (77%) of this group were verified as having voted, compared with only 36% among those who said they were "probably" going to vote. The "probably will vote" group also has a skewed partisan makeup (36% Republican, 54% Democrat) compared with those who said they would definitely be voting (46% Republican, 48% Democrat).

A different way to get at a similar concept is by asking respondents to rate their likelihood to vote on a 1 to 10 scale. Fully 75% of respondents rate the likelihood as a 9 or 10, giving them a point on the scale. A 75% majority of those voters are verified as having voted. By comparison, only one-third (34%) of those rating a 7 or 8 were verified as having voted.

Both of these measures do a good job of identifying nonvoters. Most of those who say they won't vote, or that there is a low likelihood that they will vote, in fact do not turn out. But the challenge in surveys is that relatively few registered voters say they won't vote.

Measures of voter engagement

Citizens who are more interested in politics and who have been paying attention to the campaign are presumably more likely to vote than those who are less interested. To measure engagement in the election, respondents were asked how much they had thought about the upcoming election. Fully 69% of respondents said they thought "quite a lot" or "some" about the upcoming election. To measure general interest in politics, respondents were asked how often they follow government and public affairs; 77% of respondents said they follow government affairs "most of the time" or "some of the time."

These questions are weaker at discriminating between voters and nonvoters than are the measures of vote intention. The Perry-Gallup scale assigns a point to respondents who say they have thought "quite a lot" or "some" about the election, but 55% of those who said they had thought "only a little" about the election also voted – and this group was one-fifth of the sample (21%). In addition, a substantial 43% turned out among those who said they follow government and public affairs "only now and then."

Measures of voter engagement

	Share of total	% who are verified voters	*Vote intention (pre-election)* Republican	Democratic
	%	%	%	%
Thought about election				
Quite a lot	29	82	56	39
Some	40	62	40	52
Only a little	21	55	41	51
None	10	29	29	69
	100			
Follow government and public affairs				
Most of the time	40	78	54	40
Some of the time	37	62	39	54
Only now and then	16	43	31	60
Hardly at all	8	30	43	52
	100			

Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter file. Categories in bold contribute to the likely voter index.

PEW RESEARCH CENTER

Measures of past voting behavior

Measures of past voting behavior are central to any gauge of the likelihood of voting. The Perry-Gallup index uses three measures of past voting: whether an individual voted in the previous presidential election, the individual's own assessment of how regularly he or she votes, and whether he or she has ever voted in their local precinct or election district. (Since respondents ages 18-21 may not have had the opportunity to vote in previous national elections, they typically are given an additional 1-point boost on the scale to compensate.)

Self-reported measures of past voting behavior

	Share of total	% who are verified voters	*Vote intention (pre-election)* Republican	Democratic
	%	%	%	%
2012 vote (self-reported)				
Voted	87	70	45	49
Did not vote	12	17	41	49
Too young to vote	1	27	18	73
	100			
How often vote				
Always	45	82	48	46
Nearly always	33	59	42	50
Part of the time	13	34	41	50
Seldom	9	17	28	65
	100			
Ever voted in precinct or election district				
Yes	83	70	46	48
No	17	26	36	57
	100			

Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter file. Categories in bold contribute to the likely voter index.

PEW RESEARCH CENTER

Past vote has been shown to be a strong predictor of future behavior. According to the voter file, respondents who said they voted in the 2012 presidential election were more than four times as likely as those who did not vote to have voted in 2014. While some voters who say they did not vote in 2012 still turned out in 2014 (17% were verified as voters), an analysis of respondents matched to the voter file shows that only 5% have a record of voting in 2014 but not 2012.

Not surprisingly, self-reported regularity of voting also discriminates between voters and nonvoters. Fully 78% of respondents say they always or nearly always vote in elections, and 82% of those who say they "always" vote were verified as 2014 voters, as were 59% among those who said they "nearly always" vote. That number drops significantly for those who say they vote "part of the time" and those who say they "seldom" vote – only 34% and 17%, respectively, had a verified voting record for 2014.

Measure of past voting also implicitly captures knowledge of how and where to vote. One standard question in the likely voter index asks about previous voting in the respondent's precinct or election district, and another asks about knowledge of where to vote. (This question may become increasingly irrelevant in an era of early voting and "just in time" information on smartphones.) In midterm elections, only a question about past voting in one's precinct is typically included. Fully 83% of respondents say they had voted in their precinct in the past, and 70% of these are verified as 2014 voters. Among those who said they hadn't voted in their local election venue, just 26% were verified voters in 2014.

Voter file measures of past voting behavior

	Share of total	% who are verified voters	% who are self-reported voters	*Vote intention (pre-election)* Republican	Democratic
	%	%	%	%	%
Verified past vote					
Voted in 2012	78	75	82	46	48
No record	22	21	49	37	55
	100				
Voted in 2010	55	84	90	51	45
No record	45	36	56	36	55
	100				

Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter file.

PEW RESEARCH CENTER

Voter file records

Pollsters who draw their election samples from databases of registered voters – or those who are able to match their RDD respondents with the national voter file – have an additional source of information about past voting behavior: a verified record of voting in past elections. Voting history can also be used in a likely voter prediction model in the same way as the respondents' self-reports of past voting.

As described earlier, 89% of the self-reported registered voters in the panel who completed both the pre-election and post-election waves were matched to a voter file that contains records of past voting dating back several election cycles.

More details about Pew Research Center's methodology for estimating likelihood to vote are available at http://www.people-press.org/files/2011/01/UnderstandingLikelyVoters.pdf.

Appendix B: The choice of a turnout measure

There are two indicators of voter turnout available for the type of analysis in this report: (1) each respondent's self-report in the post-election survey and (2) a voter file record of turnout. Among registered voters, 63% have a voter file record indicating that they voted in 2014 ("verified voters") and 75% said they voted ("self-reported voters").

Comparing the two approaches reveals that nearly everyone (99%) recorded in the voter file as having voted also reported that they voted. Conversely, almost no one (2%) who said that he or she did not vote has a record of having voted. Thus, a voter file record of voting almost certainly identifies actual voters. The problem is that a considerable number of respondents who said they voted do not have a corresponding record of voting: 26% of those with no voter file record of turnout say that they voted. These individuals may be correctly reporting their vote but were missed by the voter file, or they are misreporting their vote.

It is well understood that people over-report socially desirable behaviors such as voting. In our sample, 75% of our registered voter sample said they voted. That would imply a national turnout rate of about 53%, far higher than the actual rate of 36%. Some of this difference could be accounted for if politically engaged people are overrepresented in the panel. But the magnitude of this difference seems particularly large, given that the sample is already limited to self-described registered voters (94% of whom have a registration record on file) and the survey is weighted to match population demographic characteristics that are themselves strongly correlated with voter turnout. Both of these characteristics of the survey should mitigate the effects of nonresponse bias.

The verified voter turnout was 63% of registered voters (and 64% of verified registered voters), which implies a national turnout rate of about 44% – higher than documented by the total ballots counted, but less so than the rate based on self-reported turnout.

Berent, Krosnick and Lupia (2011) argue that much of the discrepancy between self-reports and voter file information is a result of errors in the matches or the voting records. More recently, Jackman and Spahn (2015b) estimate that at least 11% of the adult population is not listed on commercial voter files, and that the characteristics of those who are missing are quite different from those who are listed. Indeed, we find that the kinds of respondents who report voting but have no record of doing so are more mobile and thus more likely to be missed by the company that assembled the voter file. Among these discrepant cases, 48% have lived at their current address for five years or longer, compared with 70% among verified self-reported voters. Of course, length of tenure at an address is itself related to the likelihood of voting, so the shorter tenure of residence

of these individuals is both a reason they would be missed by the voter file and a reason they might not have voted.

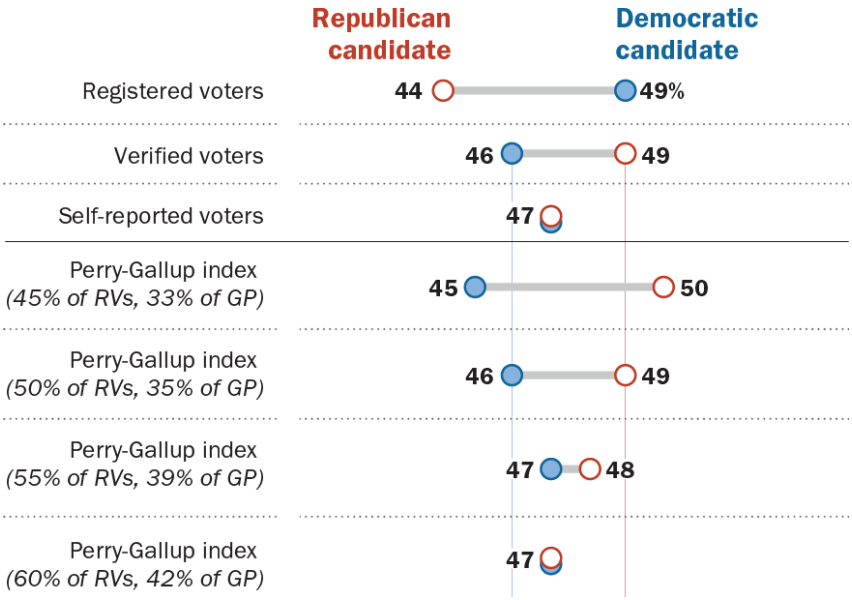
Alternatively, Ansolabehere and Hersh (2012) make the case that technological change and new legal requirements have resulted in significant improvements in the quality of the voter files. They argue that the vast majority of discrepancies result from misreporting rather than voter file errors.

Ultimately, verified vote was chosen as the "ground truth" because the error introduced by over-reporting of turnout was judged to be greater than the error resulting from mismatching. Matching errors are indeed problematic, but likely to afflict general public samples more severely than samples like the one employed in this analysis.

Appendix C: Sensitivity to the turnout forecast

The candidate preferences of voters and nonvoters in 2014 were very different. This fact makes cutoff methods very sensitive to the chosen turnout threshold. Using the Perry-Gallup method, the forecast margin ranges from a tie vote (47%-47%) with a more inclusive model (a turnout forecast of 60% of registered voters, 42% of the general public) to a 5 percentage point Republican advantage with a more restricted model (a forecast of 45% turnout among registered voters, 33% of the general public). Using the vote among verified voters as a yardstick for where voter preferences stood at that time, a turnout forecast of 50% of registered voters (35% of the general public) comes closest to the benchmark, yielding a Republican lead of 3 points (49% to 46%).

Sensitivity of Perry-Gallup index



Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter file.

PEW RESEARCH CENTER

Given that estimates for the 2014 general election put the national turnout at 35.9% of the voting eligible population, it would seem that this is the correct cutoff to use. But it is clear that the survey sample used here – registered voters who completed two waves of interviews and were matched to the voter file – overrepresented likely voters, since 63% of them (not 50%) are verified as having voted. As a result, a 60% turnout cutoff for registered voters (42% of the general population) was used for analysis.

Most survey samples are likely to have a similar bias, if not to the same extent. The problem is that it is difficult to know how much a given sample overrepresents likely voters. For this sample in this election, the Perry-Gallup method produces a forecast that is too Democratic when the appropriate cutoff is employed.