Volume 0 Issue 0
DOI: 00.000
ISSN: 2644-2353

Forecasting Turnout

Stephen Ansolabehere[†,*], Jacob R. Brown[†], Kabir Khanna[◇], Connor Halloran Phillips, Charles Stewart III
[†] Department of Government, Harvard University
[‡] Department of Political Science, Boston University
[◇] CBS Decision Desk
Department of Political Science, Vanderbilt University
Department of Political Science, Massachusetts Institute of Technology

Abstract. We evaluate the predictive power of the leading explanatory models of turnout in the existing academic literature. We compare the power of using registration, lagged vote, demographics, electoral competition, and early vote data to predict turnout. We specify models to capture each of these approaches to understanding turnout, fit those models to the relevant data from prior elections, and use the estimated parameters from prior years and the relevant observable data from the day of the election in the current year to predict that year's election.The simplest and most naive model, the Registration Model, out-performed other models in predicting 2016 turnout using 2012 election data and 2020 turnout using 2016 election data. These findings are consistent with classic understandings of which factors most drive turnout, and demonstrate that in modern elections the propensity of registered voters to turnout in Presidential elections is fairly stable. Saturated models that combine many of these predictors are common in the academic literature that attempts to explain levels of turnout. We find that such saturated models overfit the data and lead to less good predictions than parsimonious models.

Keywords: Elections, turnout, prediction, registration, modeling

Media Summary

The Media Summary should be written in plain language to highlight the key messages of the article, in ways that can be understood by the general public and cited by the media directly and accurately. It therefore should avoid technical terms or language designed for academic communications. It should not exceed 400 words, and more succinct, the better.

1. Introduction

Prediction is a basic standard against which scientific progress can be gauged. Much political science concerns itself with measurement, explanation, and causality, but pays little attention to prediction. Ultimately, though, the value of our science rests with our ability to predict. This

[*] sda@govharvard.edu

paper considers one such political science problem – predicting turnout. Is our understanding of behavior and the processes that produce it good enough to predict future behavior using a model that captures our understanding? Developing a forecasting model of turnout offers insight into how we explain participation but is also of tremendous practical value.

Our own experience with the problem of forecasting turnout comes every election night. We are members of a team of analysts at CBS News headquarters in New York. We examine the real time, but incomplete, vote tallies from the states and project as who won the US Presidency and seats in the US House and Senate based on those admittedly incomplete data. We work much like similar groups at Fox, ABC, NBC, CNN, the New York Times, and the Associated Press. When we make a determination it is based on a projection of the total votes that were cast, who won the votes that have been counted, and what the likely outcome will be among the votes that have yet to be counted. To determine correctly who won and who lost requires, first, an accurate prediction of the total votes that will be cast. In some states it will take weeks to finish counting the ballots, yet on election night we have fairly accurate predictions of turnout and of the winners. If, however, the turnout predictions are wrong, then the projection of the winners and losers may also be incorrect, as it was in 2000.[1]

The national media are not alone in making predictions about turnout. Every election administrator in the United States needs to predict how many votes are likely to be cast in order to plan and manage the election. Election lawyers and advocates for election integrity detect irregularities by comparing actual tallies to expected vote totals.[2] Campaign organizations use turnout predictions to know where to make a final push for voter mobilization.[3] And, academic researchers need to project a counter-factual level of turnout in order to measure the effects of interventions, such as campaign communications or new voting procedures (A. Fowler, 2015).[4] The problem, in a nutshell, is this. On the morning of election day, what is our best forecast for what turnout will be throughout the nation, in every state, and in each Congressional District. These are the jurisdictions for which forecasts need to be made for election night projections of who won, for monitoring election integrity, and for gauging the counterfactual level of turnout in the absence of field experiments and other interventions.

The scholarly literature on forecasting aggregate turnout is surprisingly thin. Extensive research, relying on self-reported vote and, increasingly, on data from voter files, examines which demographic groups voted and who did not (Hersh, 2015; Leighley & Nagler, 2013; Rosenstone et al., 2003). And, a long-lineage of research has sought to measure the effects of changes in election laws and in the demography of the US on aggregate turnout (Burnham, 1974; Converse, 1972; Rusk, 1974). Much recent work tests the causal mechanisms behind individual turnout decisions (Blais, 2006; Cantoni & Pons, 2022; Enos & Fowler, 2014; J. H. Fowler & Dawes, 2008; Gerber et al., 2003). Almost none of this work has used these models to predict turnout levels in subsequent elections, a task we face every election year.

This paper examines five strategies and types of models for forecasting turnout. Our approach is to specify different models, fit those models to the relevant data from prior elections, and use

[1]The models explored in this article are not the models CBS uses on election night. Rather, this article explores components of forecasting models to understand the strongest signals of turnout.

[2]See, for instance, Nick Corasaniti, "Voting Rights and the Battle Over Election Laws: What to Know," *New York Times, December 29, 2021, https://www.nytimes.com/article/voting-rights-tracker.html.*

[3]The leading discussion of these techniques is found in Green and Gerber, 2004.

[4]Perhaps the most sophisticated study to date in this long line of research is Cantoni and Pons, 2021.

the estimated parameters from prior years and the relevant observable data from the day of the election in the current year to predict that year's election. The models we examine are leading models of participation and turnout in the existing academic literature that express an important line of thinking or understanding of electoral participation. They are not exclusive of one another, and it may be possible to combine them. We treat each as a distinct approach and examine each separately. One caution in combining these models is that our own attempts to do so led to overfitting the data from prior elections, causing forecasts to err badly.

The first model is the simplest. We call it the Registration Model. Assume that the percent of registered voters who turnout in a jurisdiction is the same year-to-year; all that drives aggregate turnout is total registration. This model has its roots in Wolfinger and Rosenstone, 1980 classic *Who Votes?*. They describe a two stage process of registration and voting, and note that registration is a significant screen in the process.

The second model considers turnout as a dynamic process, where the best predictor of future voting at the individual level is whether an individual voted in the prior election. Models that predict turnout using lagged vote alone have been deployed in the individual literature to capture "habitual" behavior (Gerber et al., 2003). Although it is strange to call a behavior that one does every 2 or 4 years a habit,[5] the idea behind this approach is either that lagged vote is a sufficient statistic that summarizes the characteristics of the people in an area and their propensities to vote or that lagged vote captures a behavioral change such that people who voted become more likely to vote and people who did not become less likely to vote. In practice, lagged vote tends to underestimate turnout because it does not capture increases in the population. The best predictor of whether an individual votes is whether that individual voted in the prior election. By that reasoning, the best predictor of voting in any given area j is the turnout in that area last election.

The third model is the Demographic Model, and it represents the most widely used approach for explaining turnout at the individual level. This model posits that every demographic group has its own propensity to vote; variation and changes in the demographic composition of the electorate are the basis for projecting where and when turnout will be higher or lower. See, for example, Rosenstone et al., 2003 and Leighley and Nagler, 2013.

The fourth model is the Competition Model. It asserts that campaign activities or competitiveness of elections explains where and when turnout will be higher or lower relative to a normal or average level of turnout. Researchers have found that closeness of the election is a good short-hand for the wide range of activities and factors that affect competition; also the absence of an election, or being a non-battleground state often proxies for competition. We use closeness of the election in a state or district as the indicator of competitiveness and campaign intensity and lagged turnout to gauge expected turnout.

A fifth model uses the Early and Absentee vote. We first encountered this model through conversations with county election offices and their projections, some of whom use the early vote to project total turnout. The total vote can be separated into the total Early and Absentee vote and the total Election Day Vote. Divide registrants into those who voted absentee (which one knows on election day) and those registrants who voted on Election Day; then, use past election data to estimate the percent of registrants who voted on election day, given that they did not vote absentee, in order to estimate the number of registrants who have not voted Early and Absentee but may

[5]One doesn't smoke a cigarette every 2 years and call it a smoking habit.

vote on Election Day. This model has a particular advantage in states and counties were absentee and early voting is universal (e.g., Oregon) or the vast majority of votes cast (e.g., Arizona).

Finally, we combine all five approaches in a single "super model." In this model, one may think of lagged vote as measuring the differential vote propensity of individuals, and thus of areas, and the other factors, such as changes in registration, demographics, and competitiveness as accounting for important variation around that baseline.

One model that appears in the literature that we do not present here–the "likely voter model" of survey researchers. The Gallup Poll popularized likely voter model. Their approach was to ask people how likely they are to vote, and use those people who said they are likely to vote as the predicted turnout. These measures and estimates provided sufficiently unreliable estimates for projecting national turnout (let alone local turnout) that Gallup no longer makes a turnout projection. There is a further limitation of the survey data, which is that there are not sufficiently large samples at the CD, county, or even state level to forecast turnout at those levels using survey data.

A clear winner emerges in our analysis. To estimate and test these models we developed a database of registration, turnout, election results, and demographics at the county and CD level in the United States in 2012, 2016, and 2020. The simplest and most naive model, the Registration Model, out-performed other models in predicting 2016 turnout using 2012 election data and 2020 turnout using 2016 election data. As a check on this model, we used the Current Population Survey to estimates of turnout as a percent of registrations and data on total registrations nationally from 1972 to 2020. The simple logic suggested by Rosenstone et al., 2003 works even there. Over time, approximately the same proportion of registered voters turnout, but what changes is the number of registrants in the system. Although the aim of this exercise is not to explain turnout, the implications of this analysis for the study of participation are clear. The propensity of registered voters to turnout in Presidential elections is fairly stable. The institutions and rules that shape registration, to a first order, shape the level of turnout in the United States.

The combined model provides a humbling lesson as well. Adding more variables to the Registration Model actually made predictions worse!

2. Models and Methods

2.1. **Models.** We investigate the predictive power of five different approaches, plus a combined model.

Model 1. The Registration Model

We first consider the Registration model, which predicts turnout as a function of the number of registered voters in a county or district during an election. Let V_t be the total number of Votes cast at year t and R_t be the total number of persons registered to vote at year t. Assume that the percent of registered voters who vote is constant, v, from year to year, and that the only factor that varies is the number of people who are registered. Then,

$$V_t = vR_t$$

R_t is known in advance of election day, and the percent of registered voters who vote is observed in the prior election as $\frac{V_{t-1}}{R_{t-1}} = v$.

To the extent that the turnout rate among registered voters does not vary across years, the number of registered voters will be highly predictive of total turnout. We can estimate the relationship between total votes and total registration across counties or districts through the following linear model fit to a single year t of the data:

$$V_{t,g} = \alpha^{(t)} + \beta^{(t)} R_{t,g} + \epsilon_g^{(t)} \tag{2.1}$$

where V_t is the total number of Votes cast at year t in geography g, $alpha_t^{(t)}$ is the intercept for the model in year t, $R_{t,g}$ is the total number of persons registered to vote at year t in geography g, and $\epsilon_g^{(t)}$ is the error term. The quantity of interest is
$beta^{(t)}$, which represents the increase in total votes that is expected from an additional registered voter in the electorate in year t.

We then predict turnout in the *next* election (election $t+1$) by the following model:

$$\hat{V}_{t+1,g} = \alpha^{(t)} + \hat{\beta}^{(t)} R_{t+1,g} \tag{2.2}$$

Model 2. The Lagged Turnout (or Dynamic) Model

The lagged vote model considers whether the best predictor of individual turnout is whether that individual voted in the prior election. Likewise, the best predictor of voting in any given area j is the turnout in that area last election. For this model, we estimate the following linear model across counties or districts:

$$V_{t,g} = \alpha^{(t)} + \beta^{(t)} V_{t-1,g} + \epsilon_{t,g} \tag{2.3}$$

where $V_{t,g}$ is the number of votes in geography g during election t, $\alpha^{(t)}$ is the intercept for the fitted model from the data for election t, and $\epsilon_{t,g}$ is the error term. The quantity of interest in this case, $beta^{(t)}$, represents the correspondence between past turnout and future turnout across geographies. We estimate these parameters from this model and then predict turnout in the following election as so:

$$\hat{V}_{t+1,g} = \hat{\alpha}^{(t)} + \hat{\beta}^{(t)} V_{t,g} \tag{2.4}$$

Note that the Registration and Lagged Vote Models can be combined. In the combined model, we can think of the lagged vote as measuring the differential vote propensity, while the simple Registration model assumes that propensity is fixed. In the results section we present performance metrics for each model separately and then demonstrate the performance of different combinations of our five turnout models.

Model 3. The Demographic Model

Next we consider the predictive power of county or district-level demographics. Let X_t be a set of demographic characteristics. Consider turnout as a function of these demographics prior to an election, where each demographic variable has an additive predictive effect on total votes in that

election. Thus,

$$V_t = X_t\beta + \epsilon_t$$

Assuming the coefficients β are fairly stable, then the vote can be forecast by estimating β from the prior election and using the estimated coefficients times the current values of X_t to forecast V_t. As such, we estimate the following model across counties or districts for a prior election t to estimate the vector of coefficients β which consists of the coefficients for each demographic variable:

$$V_{g,t} = \alpha^{(t)} + \beta^{(t)}\mathbf{X}_{g,t} + \epsilon_g^{(t)} \tag{2.5}$$

With the estimated coefficients from this model we then predict turnout in the next election through the following model:

$$\hat{V}_{g,t+1} = \hat{\alpha}^{(t)} + \hat{\beta}^{(t)}\mathbf{X}_{g,t+1} \tag{2.6}$$

Model 4. The Competition Model

The fourth model we consider conceptualizes turnout as a function of the electoral competition of a given election and/or area. Let $C_{g,t}$ be the competitiveness of the election in geography g during election t, measured by closeness in the polls, campaign spending, or closeness of past election returns. Let $X_{g,t}$ be the total population in each county or district as of election year t.[6]

$$V_{g,t} = \alpha^{(t)} + \beta^{(t)}C_{g,t} + X_{g,t} + \epsilon_g^{(t)} \tag{2.7}$$

We predict turnout in the next election through the following model using the coefficients from equation 7:

$$\hat{V}_{g,t+1} = \hat{\alpha}^{(t)} + \hat{\beta}^{(t)}C_{g,t+1} + X_{g,t+1} \tag{2.8}$$

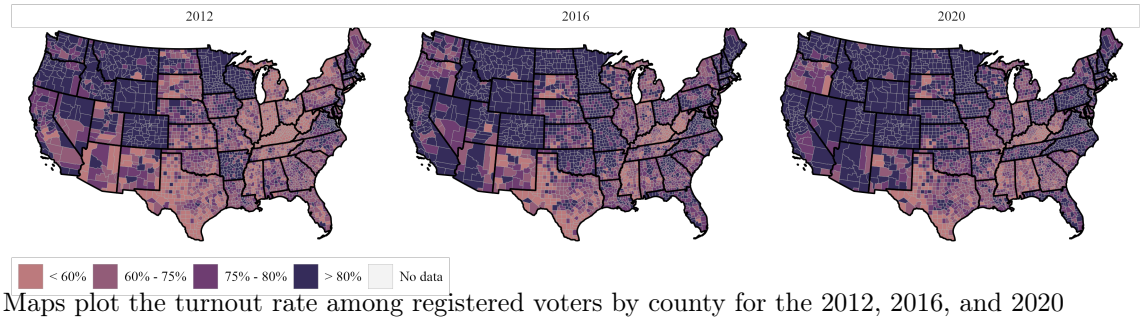
Model 5. Early Vote Model.

Define D_t to be the total votes cast on Election Day and A_t to be the total Advanced Vote. Let R_t be the total number of registered voters in election t. We can conceptualize the total number of votes cast as a function of the early voting rate and the election day turnout rate. The question for this model is how early information about the advanced voting rate is predictive of higher total turnout, both mechanically so (as advanced votes count towards total turnout), and as a barometer that election day voting may be higher in the election (i.e. evidence of election intensity). Consider the following model:

$$V_t = R_t[\frac{D_t}{R_t - A_t}\frac{R_t - A_t}{R_t} + \frac{A_t}{R_t}] = R_t[d_t(1 - a_t) + a_t]$$

[6]We include county or district total population in the estimation of the Demographic and Competition models, since the core variables in those models are proportions, while the outcome in all models is total votes in a county or district.

Figure 1. Turnout among registered voters across counties, 2012-2020



Maps plot the turnout rate among registered voters by county for the 2012, 2016, and 2020 elections.

$a_t = A_t/R_t$ and $d_t = D_t/(R_t - A_t)$. The variables R_t and a_t are observed; measure d_t using the prior election.

In practice, we estimate this model similar to the previous models. First estimating the relationship between advanced vote and total vote in a prior election then fitting that to the next election. We also demonstrate the capacity to combine advanced vote information with other models (particularly Registration and Lagged Vote Models).

$$V_{g,t} = \alpha^{(t)} + \beta^{(t)} A_{g,t} + \epsilon_g^{(t)} \tag{2.9}$$

We then predict turnout in the next election through the following model:

$$\hat{V}_{g,t+1} = \hat{\alpha}^{(t)} + \hat{\beta}^{(t)} A_{g,t+1} \tag{2.10}$$

2.2. **Data.** To estimate each of these models, we developed a database of registration, turnout, election results, and demographics at the county and CD level in the United States in 2012, 2016, and 2020. These data were provided by CBS Decision Desk and MIT Election Lab, as supplemented with election results data collected directly from states. The data consist of county and Congressional House district election results for presidential elections 2012, 2016, and 2020. For each county and district, we observe the total number of registered voters in that geographic unit on election day, the total number of votes cast, presidential vote share, and Congressional vote share. We combine these data with data from the 2016 and 2020 Election Administration and Voting Survey, which provides information on total advanced vote as of election day for the 2016 and 2020 elections.[7] Figure 1 shows the turnout rate among registered voters by county for the 2012, 2016, and 2020 elections.

We further combine these data with information from the 2006-2010, the 2010-2014, and the 2014-2018 American Community Survey (ACS) from the United States Census. From these data we source the demographic variables used in the estimation of the Demographic Model. We use the

[7]Some counties and districts in the data have vote totals that exceed the total number of registrants. These discrepancies are due to registration counts being current as of the weeks immediately prior to election day for each year, and these cases are disproportionately in states with Election day registration. We do not correct these cases in our data, since forecasters of an upcoming election must contend with such data limitations.

2006-2010 ACS to measure demographic variables for the 2012 presidential election, the 2010-2014 ACS for the 2016 presidential election, and the 2014-2018 ACS for the 2020 presidential election. We do use these earlier ACS data for each election since at the time of any given election the more recent (i.e. 2015 or 2016 ACS information for the 2016 election) is either not yet released or is still being collected. The demographic variables we use from the ACS include total county and district population, the percentage of the population in a county or district that is married, graduated from college, the median household income, the percentage of the population that is white, black, and hispanic, and the percentage of the population in each of four age categories: 15-24, 25-34, 35-64, and above 65 years of age.

2.3. **Estimation and Estimates.** To estimate the Registration model, we first subset our data by state, and estimate separate models within each state. For each state, we estimate regressions in the form of equation 1 on the 2012 results and the 2016 election results, modeling turnout totals in each election as a function of registration totals in each election. We do so for both counties and districts, separately, as the units of analysis. We then apply the coefficients from these regression models to predict turnout in the next presidential election. So with the 2012 regression parameters we predict 2016 turnout using 2016 registration, and the 2016 regression parameters we predict 2020 turnout using 2020 registration.

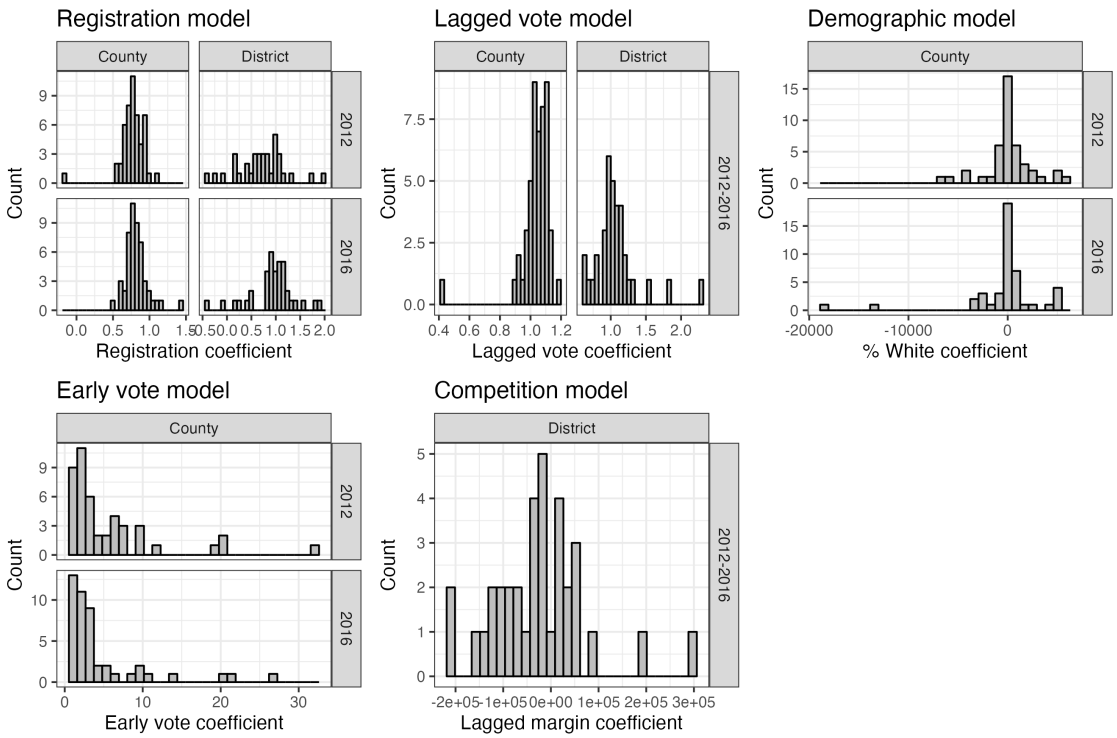
To estimate the Lagged Vote model, we conduct a similar exercise, except that we fit regressions for each state measuring the relationships between 2012 turnout and 2016 turnout, and then use those parameters to predict 2020 turnout using 2016 turnout.

We estimate the Demographics and the Early Vote models following a similar approach to the Registration model. We use regression analyses using data for data from 2012 to estimate the parameters of these models. We, then, predict 2016 turnout. Likewise, we perform regressions for the 2016 data and use that to predict 2020 turnout. The Early Vote model is only estimated on counties, since EAVS does not have early vote data aggregated by Congressional districts.

To estimate the Competition model, which uses lagged vote margin, we use 2012-2016 data to estimate regressions we then use to predict on test data from 2016-2020. The Competition model is only estimated at the district level, since competition is operationalized through lagged vote margin of previous House of Representative elections at the district level. Our main results use weighted ordinary least squares regressions, using the number of registered voters in a county or geography as weights. Unweighted results are very similar to the weighted results.

Figure 2 plots the distribution of coefficients across models. For both the county models and the district models, and across years, the registration coefficients are clustered around 0.8 and 0.9. respectively, ranging from -0.139 to 1.46 for counties and -0.442 to 6.59 for districts. This means, that on average across states and models, each additional registrant in a county or district translates into 0.8-.0.9 additional votes in that county/district in that election. The lagged vote coefficients range from 0.429 to 1.19 for counties (0.608 to 7.63 for districts), and are generally clustered around 1, indicative of the strong likelihood that a voter in a past election will vote in the next election. The Demographic model yields many coefficients, so Figure 2 only shows the distribution of coefficients for percent White. These coefficients have a wide range, but on average are not predictive of increased or decreased turnout in a county or district, net of other variables in the Demographics model. The early vote coefficients range from 0.83 to 31.84, but are generally concentrated between 0.83 and 5, demonstrating a strong positive relationship between early vote counts and total vote counts. This positive correlation is in part mechanical, since each additional early vote equates

Figure 2. Coefficients from state subsets across models



to an additional total vote. But the extent this coefficient varies in size is a function of whether early vote is a substitution for Election day turnout or a sign of heightened participatory interest in the election. The coefficients for the Competition model, lagged vote margin at the district level (ranging from 0 to 1), are centered close 0 but range widely.

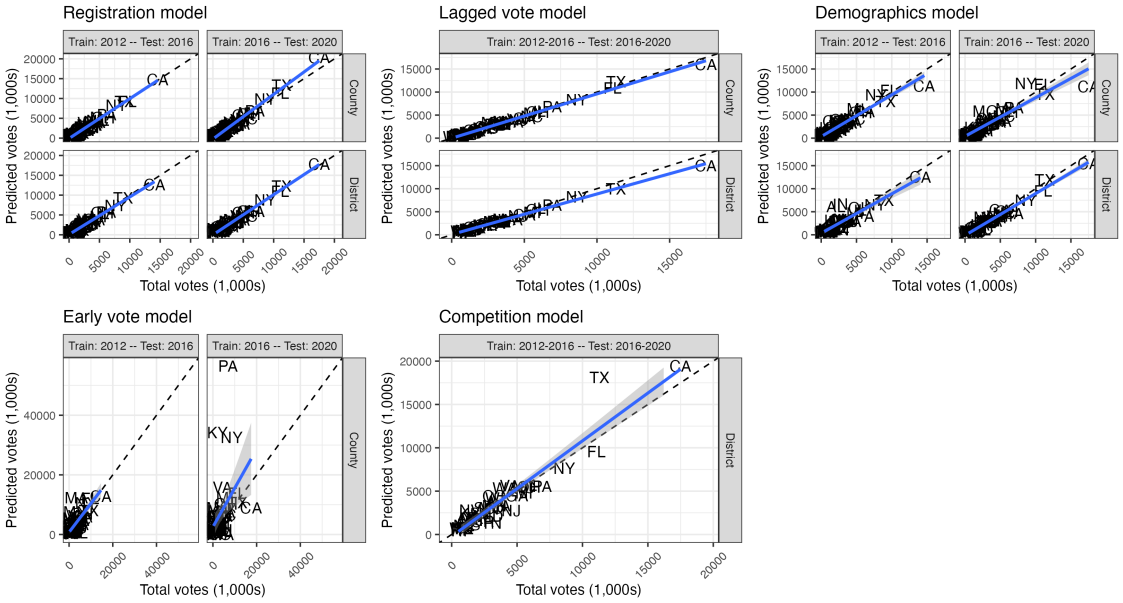
3. Comparison of Models

We compare the predictive performance of the models using the bias and root mean squared errors of the predictions. Figure 3 plots state level predictive vote counts against actual vote counts for each model (Registration, Lagged Vote, Demographics, Early Vote, and Competition) fore relevant election years and across counties and districts.

The scatter plots show that the registration model predicts state vote totals across years and geographies well. It slightly understates 2016 turnout (using 2012 parameters) and slightly overstates 2020 turnout (using 2016 parameters). We are particularly surprised by the performance of the model in 2020 given the substantial changes in the electoral procedures in the states and the challenges people faced voting during the COVID-19 pandemic.

The Lagged Vote model tends performed somewhat less well than the Registration Model. Lagged Vote tends to understate turnout, but overall the predicted and actual state totals show high correspondence.

Figure 3. Predicted votes against total votes across models



Similarly, the Demographics Model shows high correspondence between the actual and observed vote, but a tendency to understate predicted turnout. Lagged Vote and Demographics fail to capture upward trending in turnout.

The Early vote model does not perform well. It dramatically overpredicts turnout in many states, and does a particularly poor job in Pennsylvania, Kentucky, and New York in the 2020 elections.

The Competition model on average overstates turnout but overall performs well, with most predicted and actual totals close in most states. It has much larger deviations in the largest states, such as Texas and California.

It is also worth noting that the models perform much better when data are measured at the Congressional District level, rather than the County level. This is important because elections are administered at the county level, and much of the data reporting and analyses use the county as the level of analysis. We return to the question of why these differences arise at the end of this paper.

The scatter plots provide visual evidence of the predictive performance of each model. To more formally assess and compare this performance, we calculate the overall bias and root mean squared error for each model (separately for appropriate time periods and geographic units of analysis). Table 1 reports these performance statistics for the Registration and Lagged Vote models.[8] We report bias and RMSE in total vote counts and as a percentage of average county or district level vote totals. The Registration model at county level predicts 2016 turnout using 2012 training data with a bias of -2.4%, and predicts 2020 turnout using 2016 training data with a bias of 5.6%. The RMSE for the 2016 turnout prediction is 42%, while the RMSE for the 2020 turnout prediction

[8]The district analyses exclude the 7 states with only 1 district (Alaska, Delaware, Montana, North Dakota, South Dakota, Vermont, and Wyoming). The 2012-2016 district analyses also exclude Virginia, North Carolina and Florida due to data availability.

Table 1. Evaluation of Registration and Lagged Vote Models

	Registration				Lagged Vote	
	County		District		County	District
Training	2012	2016	2012	2016	2012-2016	2012-2016
Test	2016	2020	2016	2020	2016-2020	2016-2020
Bias (Total)	-1,065	2,803	-6,884	3,058	-2,171	-35,275
Bias (% of Avg. Vote)	-2.4%	5.6%	-2.2%	0.8%	-4.3%	-9.7%
RMSE (Total)	18,436	14,564	34,589	43,574	11,983	42,450
RMSE (% of Avg. Vote)	42%	29.1%	11.1%	12%	23.9%	11.7%
Units (Train)	3,151	3,151	377	428	3,151	377
States (Train)	50	50	40	43	50	40
Units (Test)	3,151	3,151	377	428	3,151	377
States (Test)	50	50	40	43	50	40

is 29.1%.The district models show improved performance, with biases of -2.2% (2016 turnout) and 0.8% (2020 turnout). The RMSE for these estimates are 11.1% (2016 turnout) and 12% (2020 turnout). The Lagged Vote model exhibits lower variance, but overall greater absolute bias than the Registration model. Predicting 2020 turnout using 2012-2016 training data at the county-level with the Lagged Vote model has a bias of -4.3% and an RMSE of 23.9%. The district-level prediction performs worse, with a bias of -9.7% but an RMSE of just 11.7%.

Table 2 presents the performance metrics for the Demographics model, Early Vote model, and the Competition Model.[9] The Demographic model at the county level offers similar performance to the Registration model in terms of bias, but suffers from higher root mean squared error. Predicting 2016 county turnout using 2012 training data using the Demographic model results in a bias of 2.7% and an RMSE of 60%. For the 2020 county turnout prediction, the bias is 4% and the RMSE is 68.8%. The district predictions are more uneven, with the 2016 turnout prediction performing very well – with a bias of -0.1%, although an RMSE of 43.6% – but the 2020 prediction performs poorly, showing a bias of -13.2% (although its RMSE is just 24.5%). Overall, the Demographics model can predict turnout well, but is more uneven across elections in its performance than the Registration model. Further, the RMSE from the Demographics model is always higher than that for the Registration model, so while biases of the two models may be similar at times, the Demographics Model often offers higher uncertainty in its forecasts.

The Early Vote model has very high bias and RMSE. Predicting 2016 turnout using 2012 training data at the county level with the Early Vote model has a bias of 24.7% and an RMSE of 292.9%. For the 2020 turnout prediction, these metrics are even worse, with a bias of 110.3% and an RMSE of 633%. Thus, these predictions show both high bias and uncertainty, and Early Vote is not a reliable predictor on its own of total vote counts. The Competition model does much better, with a bias of -1.6% for predicting 2020 turnout using 2012-2016 training data, which is almost as low as the equivalent district-level bias (0.8%) for predicting 2020 turnout with the Registration model.

[9]The county estimates from the Demographic model exclude Alaska due to data availability.

Table 2. Evaluation of Demographics, Early Vote and Competition Models

	Demographics				Early Vote		Competition
	County		District		County		District
Training	2012	2016	2012	2016	2012	2016	2012-2016
Test	2016	2020	2016	2020	2016	2020	2016-2020
Bias (Total)	1,181	-1,985	-255	-47,882	10,839	55,286	-5,636
Bias (% of Avg. Vote)	2.7%	-4%	-0.1%	-13.2%	24.7%	110.3%	-1.6%
RMSE (Total)	26,349	34,487	135,841	88,773	128,697	317,304	262,075
RMSE (% of Avg. Vote)	60%	68.8%	43.6%	24.5%	292.9%	633%	72.2%
Units (Train)	3,110	3,110	365	426	2,883	2,873	425
States (Train)	49	49	40	43	45	46	43
Units (Test)	3,110	3,110	365	426	2,865	2,936	428
States (Test)	49	49	40	43	45	46	43

However, the Competition model does show a higher RMSE than the registration model, with an RMSE of 72.2%.

The evaluation of each individual model indicates that the Registration model is the most consistently accurate of the models. The Demographic and Competition models are competitive with the Registration Model in terms of minimizing bias, but those models show higher variance and thus larger RMSE than the Registration model. As a result, the standard error of predictions of turnout forecasts based on Demographics or on Electoral Competition will always be larger than predictions that rely on Registration rates. For election administrators, this would mean higher cost in terms of the amount of ballots, precincts, and the like that need to be prepared in order to guard against a very high turnout election in some areas. For analysts on election night, it means a much slower rate at which models will converge on a forecast on which a final call might be made. For academic researchers comparing an observed election outcome versus a counter factual, it means greater uncertainty about what that counter factual might be, and lower statistical power (and more errors in academic judgment) about the effects of innovations in election administration.

The final model examined combines these approaches. Political scientists and sociologists studying turnout frequently combine many, if not all, of the approaches into a single analysis. Often the effort is to conduct an horserace to find the most powerful explanatory account or to attribute portions of the explained variance to each of the different accounts (e.g,. Rosenstone and Hansen 1993). Rosenstone and Hansen (1993) expressly use a regression model that combines these many approach in a single cross-sectional analysis to explain changes in the level of turnout in the US from 1960 to 1990.

How well does such an approach work for our problem? We take the Registration Model as the baseline model and then added each of the others (one at at time) to it. We also combined all of the models into one “super model." Thus, we first assess four additional models, the first combining registration and lagged vote covariates in the same regression, the second combining registration and demographic variables, the third using registration and early vote, and the fourth with registration and lagged vote margin. For each combined model, we estimate it in a similar fashion as the individual models: subsetting the data by state and running county unit or district

unit models for each state across relevant time periods, predicting 2016 turnout using 2012 training data and 2020 turnout using 2016 training data where appropriate. Finally, we put all of the models into one big stew. The bias and RMSE of each of these combined models are reported in Table 3.

When we combine the Registration model with the Lagged Vote model, we predict 2020 county turnout using 2012-2016 training data with a bias of -10.1% and an RMSE of 24.1%. District turnout in 2020 with this combined model is predicted with a bias of -4.6% and an RMSE of 66.4%. The county and district turnout models show greater absolute bias than the stand alone registration model, although the RMSE is similar. Thus, there seems to be no prediction gains from combining election year registration information with lagged vote information.

Combining the Registration and Demographics models results in low bias for the county-level models in predicting both 2016 and 2020 turnout (-4.1% and -1.7% bias, respectively). This combined model outperforms the stand alone Registration and Demographic models in terms of absolute bias in predicting 2020 county turnout but not 2016 county turnout. The RMSE for the combined model county estimates are 54.9% and 58.4%. Thus, adding these additional variables increased the RMSE, relative to the stand alone Registration county models.

Worse still are the forecasts conducted using Registration plus Demographics using data at the Congressional District level. The biases for these models leap up to the double digits in predicting 2016 and 2020 turnout. The RMSE for these district estimates are 287% in predicting 2016 turnout and 26.5% in predicting 2020 turnout.

The combined Registration and Early Vote models perform well in predicting 2016 county turnout using 2012 training data (-1.4% bias and 33.9% RMSE) but perform poorly when predicting 2020 county turnout using 2016 training data (16.6% bias and 150.1% RMSE). The combined Registration and Competition models[10] performs perhaps the strongest of any model, stand alone or combined. This combined model predicts 2020 district turnout using 2012-2016 training data with a bias of just -0.5% and an RMSE of 10.2%. This represents a reduction in bias and RMSE compared to the stand alone Registration model and a reduction compared to the stand alone Competition model.The success of this model highlights the potential for registration information and electoral competition information to provide complimentary information in forecasting turnout.

Adding additional variables to the basic Registration Model, then, rarely made matters better. Perhaps the best we can say is that in some instances, such as Registration and Competition, the prediction RMSE did not get noticeably worse. In part this is because the Registration Model is itself a fairly good forecasting model, so it is challenging to improve on it. The bigger lesson is that complex models that combine many different approaches and elements are often overfitting the data. This can only be evident when looking at the predictive value of the analyses, rather than simply the fit of the models within samples. We think this is a very important lesson not only for the forecasting literature but for the entire empirical enterprise that attempts to explain why people vote.

[10]When combining the Registration and Competition models, we omit district population from the regressions, although it is included in the stand alone Competition model as stabilizing covariate since the model is predicting total votes and lagged vote margin is a proportion. The inclusion of registration counts in the combined model performs the same role.

Table 3. Evaluation of Registration Model Combined with Other Models

	+ Lagged Vote		+ Demographics				+ Early Vote		+ Competition
Registration	County	District	County		District		County		District
Training	2012-2016	2012-2016	2012	2016	2012	2016	2012	2016	2012-2016
Test	2016-2020	2016-2020	2016	2020	2016	2020	2016	2020	2016-2020
Bias (Total)	-1,189	-26,527	-1,780	-833	57,451	-40,433	-621	8,344	-1,830
Bias (% of Avg. Vote)	-2.4%	-7.3%	-4.1%	-1.7%	18.5%	-11.1%	-1.4%	16.6%	-0.5%
RMSE (Total)	12,003	41,032	24,105	29,273	893,474	96,349	14,895	75,227	36,922
RMSE (% of Avg. Vote)	23.9%	11.3%	54.9%	58.4%	287%	26.5%	33.9%	150.1%	10.2%
Units (Train)	3,151	377	3,110	3,110	365	426	2,919	2,873	428
States (Train)	50	40	49	49	40	43	46	46	43
Units (Test)	3,151	367	3,111	3,111	377	428	2,865	2,936	418
States (Test)	50	40	49	49	40	43	46	46	43

Table 4. Evaluation of All Models Combined

	+ Lagged Vote + Demographics + Early Vote	+ Lagged Vote + Demographics + Competition
Registration	County	District
Training	2012-2016	2012-2016
Test	2016-2020	2016-2020
Bias (Total)	-2,311	-36,322
Bias (% of Avg. Vote)	-4.6%	-10%
RMSE (Total)	33,308	87,272
RMSE (% of Avg. Vote)	66.4%	24%
Units (Train)	2,873	375
States (Train)	46	40
Units (Test)	2,936	377
States (Test)	46	40

4. Extensions

Two aspect of this analysis deserve further immediate attention. First, we can validate this analysis by examining a simple prediction implied by the results here. The prediction is this. Fluctuations in aggregate turnout over time at the national level ought to follow the Registration Model. Second, there is an internal puzzle that deserves further discussion: why are predictions based on congressional district level data better than predictions based on county level data?

4.1. **The Registration Model Over Time Using CPS.** One way to test the value of the Registration model is to see how well it explains trends in turnout rates over time. As is well-known among political scientists, turnout in the United States fell from 1960 to 2000, and it has increased since 2000. Does the Registration model explain long-term patterns in turnout?

There is no national measure of total registered voters in the United States based on voter files, because national voter files became available only in 2008. Research on historical registration rates relies, instead, on the Current Population Study (CPS) data from 1970 to the present. Below are the CPS estimates of registration and turnout among citizens and turnout among registered voters. Several scholars have criticized these data as being slightly inaccurate, particularly for some demographics.[11] Although not perfect these are the best available data on registration and turnout over the past 50 years. We use the estimates as reported by CPS for each year. It may be possible to improve the estimates by applying different weights,[12] but we take the data as given and assume that, even if it is off by a couple of percentage points in the levels, that it reflects trends.

	Election Cycle	
CPS Category	Presidential (Average and SD)	Midterm (Average and SD)
Share of Citizens Who Report Being Registered to Vote	71.9 (1.6)	67.0 (1.3)
Share of Registered People Who Report That They Voted	87.7 (2.5)	71.6 (4.0)
Share of Citizens Who Report Report That They Voted	63.1 (2.8)	48.0 (3.2)

Table 5. Registration and Turnout 1970-2020, Current Population Survey

We apply the model to the CPS data at the national level. For each year, we calculate the predicted turnout as the registration rate times the turnout rate of registered voters in the prior presidential year, for presidential elections, or the prior midterm year, for midterm elections. For instance, the predicted turnout in 2020 equals the voting age population times the CPS estimate of turnout as a percent of registered voters in 2016 and multiply that times the registered percent in 2020.

$$T_1 = (T_0/R_0)R_1$$

The model does surprisingly well. At least, its performance certainly surprised us. We calculated the percentage error of the estimated total turnout: the actual turnout in a year minus the model's predicted turnout, divided by the actual turnout. The percentage error averages -.0009, with a standard deviation of .059. That is, on average there is no bias in the Registration Model's predicted turnout over time, and its mean squared error is about 6 percent. According to the CPS report on "Voting and Registration" for 2022, the standard error on the survey estimate is approximately

[11] In response to these critiques, the Census Bureau has performed their own evaluation of the accuracy of the CPS data. See Kurt Bauman, "How Well Does the Current Population Survey Measure the Composition of the US Voting Population?" U.S. Census Bureau SEHSD Working Paper 2018-25, July 6, 2018. https://www.census.gov/content/dam/Census/library/working-papers/2018/demo/SEHSD-WP2018-25.pdf

[12] The weights are available at: https://cran.r-project.org/web/packages/cpsvote/vignettes/voting.html

1 percent.[13] The survey standard error is independent of the model standard error, so the model standard error is approximately 5 percent and the survey standard error adds another 1 percent.

The model misses substantially in two years, 1974 and 2018. Both are midterm elections. The percent error is -12 percent in 1974 and + 18 percent in 2018. These could reflect shifts in the electorate, or changes in the survey, as the CPS implemented the correction suggested by Hur and Achen in 2016 and 2018.

The mean squared error reveals that there is excess error due to the model. Among the 13 presidential elections analyzed, the average error is -.0003 with a standard deviation of .043, substantially lower than the entire sample of observations. If one assumes independent predicted values, the expected variation is approximately 1 percent. The standard deviation of the 13 presidential predictions is .32; hence, the standard error is .01 (i.e., $.032/\sqrt{13}$). This is an apples-to-apples comparison as the turnout predictions for presidential elections are based on lagged presidential election years. The observed mean squared error of .04 is larger, and suggests that modeling error explains much of the inaccuracy of the model. The modeling errors, however, are not systematically over or underestimating turnout. Similarly, Among the 13 midterm elections analyzed, the average error is -.0016 with a standard deviation of .074, much larger than the prediction errors for the presidential elections. Among the midterm elections, the variance of the predicted value is .039 in the 13 midterm years, which implies a standard error on the forecast of .011. The observed variance of the deviation of forecasts from the true values is .07 in midterm elections.

It is unclear to us whether further improvements are possible at this level of analysis. Although the standard error exceeds sampling error, it is unclear refinements could reduce that error further. The Registration Model is very parsimonious, and only uses the degree of freedom lost by estimating turnout rates from lagged observations. Other corrections, such as for demographics, would use more information and thus more degrees of freedom. Our analysis of the CD-level data in 2016 and 2020 suggests that combining demographics with the Registration Model led to overfitting and actually worsened prediction accuracy.

The Registration Model shows no evidence of bias in its predictions about the national trends in total turnout since 1970. This fact carries an interesting substantive result. The long-term drop in turnout from 1960 to 2000 and the rebound that has occurred in the 21st Century may be accounted for by fluctuations in voter registration alone. In the late 1990s the United States implemented the National Voter Registration Act (NVRA), and in the early 2000s the parties began aggressive voter mobilization campaigns. It is plausible that the legal innovations of the NVRA provided the incentives for people to get registered and for campaign organizations to facilitate that. The result may have been a steady increase in registration, which translated into increased turnout rates.

4.2. **Counties versus CDs and the Problem of Heterogeneity.** One puzzle resulting from our analysis is the difference between the models fitted using congressional district-level (CD-level) data and county-level data. Most research and analytics uses county-level data. Most of the analytic tools used by survey research firms and media organizations rely on county-level data and estimates. Why are the predictions using the county-level data worse than those derived using the CD-level data?

The sheer number of units would lead one to think that parameter estimates based on counties should be superior. There are over 3,100 counties, but only 435 Congressional Districts. Yet,

[13]See Table 1 of https://www.census.gov/content/dam/Census/library/publications/2022/demo/p20-585.pdf

consider the estimates in Table 1. The biases for the Registration Model and the RMSE are lower using the CD-level data than using the county-level data. The larger number of counties does *not* lead to higher precision in the prediction estimator.

How is that possible? We conjecture that the problem arises from across-unit heterogeneity. CDs are all the same size, roughly 700,000 people. Counties vary considerably, from a few hundred (e.g., Loving County, Texas, Arthur County, Nebraska, and Petroleum County, Montana) to nearly 10 million people in Los Angeles County, CA. Los Angeles County, CA, has a total population equal to the combined population of the 9 least populous states. It is this variation in population size that seems to be the problem. More correctly, the turnout rate of registered voters varies randomly, but because of the extremely different unit sizes the variation becomes correlated with unit size. This is not readily corrected by simply re-weighting the data by population, as the error in the county appears to be correlated in a non-ignorable way.

That problem is less present in the CD-level data. Even though there are many fewer CDs than counties, their populations are, as a matter of law, nearly equal. The heterogeneity in turnout rates among registered voters cannot be correlated with unit population size in the CD-level data, because the CD populations are all virtually the same.

5. Conclusion

The narrow implication of this analysis is that forecasting turnout boils down to understanding the number of people who are on the voter rolls, and thus eligible to vote, and the turnout rate among registered voters. Those two factors, combined, yield a parsimonious model and accurate predictor of turnout at the CD, state, and national level. Other models, such as those that rely on demographics, lagged vote, or competition, or combinations of models, tend to perform less well in the sense that they have higher prediction root mean squared error.

The broader lesson, though, is that prediction should inform the more general enterprise of social science analysis. Too often, political scientists, sociologists, economists, and other social scientists, rely on model fit within a particular sample to test theories and analyze their empirical implications. That practice, as shown in the case of turnout, leads to overfitting. A model might seem compelling within the data under study, but when we use it to look forward we quickly find that it does not yield particularly useful predictions. Often researchers have loaded regression analyses with many variables. That is especially true when the researchers are attempting to run a horserace among competing models or approaches to understanding a problem. The results here should give all researchers pause about that particular approach to inquiry.

The approach we recommend is to use explanatory models in tandem with predictive analyses. In the analysis of turnout, that approach lands us back with one of the early, significant results in the literature: Wolfinger and Rosenstone's finding that registration was the key to explaining turnout. We find that is not only true in an analysis of what factors correlate most strongly with turnout, but which model yields the best predictions of future turnout. The results here also offer a cautionary tale. Political scientists have likely leaned too hard on their data when estimating highly saturated regression models. We should give such analyses greater weight only when they can demonstrate their predictive power, as well as the strength of correlations or causal effects.

References

Blais, A. (2006). What affects voter turnout? *Annual Review of Politicla Science*, *9*, 111–125.

Burnham, W. D. (1974). Theory and voting research: Some reflections on converse's "change in the american electorate". *American Political Science Review*, *68*(3), 1002–1023. https://doi.org/10.2307/1959143

Cantoni, E., & Pons, V. (2021). Strict id laws don't stop voters: Evidence from a u.s. nationwide panel, 2008–2018. *Quarterly Journal of Economics*.

Cantoni, E., & Pons, V. (2022). Does context outweigh individual characteristics in driving voting behavior? evidence from relocations within the u.s. *American Economic Review*.

Converse, P. E. (1972). Change in the american electorate. *The human meaning of social change*, 263–337.

Enos, R. D., & Fowler, A. (2014). Pivotality and turnout: Evidence from a field experiment in the aftermath of a tied election. *Political Science Research and Methods*, *2*(02), 309–319.

Fowler, A. (2015). Regular voters, marginal voters and the electoral effects of turnout. *Political Science Research and Methods*, *3*(02), e1–e15.

Fowler, J. H., & Dawes, C. T. (2008). Two genes predict voter turnout. *The Journal of Politics*, *70*(3), 579–594.

Gerber, A. S., Green, D. P., & Shachar, R. (2003). Voting may be habit-forming: Evidence from a randomized field experiment. *American Journal of Political Science*, *47*(3), 540–550.

Green, D. P., & Gerber, A. S. (2004). *Get out the vote!: How to increase voter turnout*. Brookings Institution Press.

Hersh, E. (2015). *Hacking the electorate: How campaigns perceive voters*. Cambridge University Press.

Leighley, J., & Nagler, J. (2013). *Who votes now?: Demographics, issues, inequality, and turnout in the united states*. Princeton University Press. https://books.google.com/books?id=iGmYDwAAQBAJ

Rosenstone, S., Hansen, J., & Reeves, K. (2003). *Mobilization, participation, and democracy in america*. Longman. https://books.google.com/books?id=QicLAAAACAAJ

Rusk, J. G. (1974). Comment: The american electoral universe: Speculation and evidence. *American Political Science Review*, *68*(3), 1028–1049. https://doi.org/10.2307/1959145

Wolfinger, R., & Rosenstone, S. (1980). *Who votes?* Yale University Press. https://books.google.com/books?id=XmspYgn-syYC