



# Division Does Not Imply Predictability: Demographics Continue to Reveal Little About Voting and Partisanship

Seo-young Silvia Kim<sup>1</sup> · Jan Zilinsky<sup>2</sup>

Accepted: 1 August 2022 / Published online: 20 August 2022  
© The Author(s) 2022

## Abstract

What are the political consequences of ongoing social sorting? We evaluate the degree of social sorting and mass polarization using the predictability of partisanship and voting decisions as quantities of interest. Contrary to expectations, demographic sorting has not produced a very predictable electorate. Models trained on nothing more than demographic labels from public opinion surveys (1952–2020) predict only 63.9% of two-party vote choices and 63.4% of partisan IDs correctly out-of-sample—whether they be based on logistic regressions or tree-based machine learning models. Moreover, demographics’ predictive power over vote choice or partisan affiliation shows a surprising stability over time. We argue that while select demographics’ marginal effects may appear to be evidence of social sorting, the joint predictability of political behavior using only demographic characteristics has been, and still is, modest at best.

**Keywords** Vote Choice · Elections · Polarization · Sorting · Ideology · Demographics · Machine Learning · Tree-based Models · Random Forests

## Introduction

Understanding whether political choices and identities are closely linked to voters’ observable characteristics is important for both theoretical and practical reasons. Theoretically, whether a readily perceivable demographic group is a reliable political base or a swing voting bloc has substantial implications for representation (Lazarsfeld et al., 1944; Panagopoulos, 2015, 2017, 2020). Moreover, if group identities tied to partisanship intensify, group divisions will increase, fueling affective polarization (Levendusky, 2018). Practically, candidates want to know to what

---

✉ Jan Zilinsky  
jan.zilinsky@tum.de

<sup>1</sup> American University, Washington, DC, USA

<sup>2</sup> TUM School of Social Sciences and Technology, Munich, Germany

extent any constituency might be persuadable (Axelrod, 1972; Fenno, 1978; Hersh, 2015).

This paper investigates the following questions: how reliable are *demographic* labels in predicting presidential vote choices and partisan identities? Moreover, do demographic groups increasingly vote (and affiliate themselves with parties) in predictable ways, exacerbating existing cleavages? We argue that the joint, out-of-sample predictive power of voters' observable demographic characteristics to infer their voting decisions and partisan affiliation provide a useful statistic to evaluate the degree of social sorting. If demographic groups have sorted themselves into political camps over the years, their behavior should be increasingly predictable.

Using public opinion surveys from 1952 to 2020, we find that predictions based on five commonly used demographic variables—age, gender, race, education, and income—have low accuracy, which is stable over time. Only 63.9% of two-party presidential vote choices and 63.4% of binary party IDs are correctly predicted in hold-out sets of respondents whose data is withheld in the model-training stage. The same is true if a broader set of demographic labels are used, including religion, working status, and marital status: the corresponding accuracies are respectively 67.4% and 67.2% in an extended model for vote choice and party ID.

Not only is the informativeness of demographic labels surprisingly muted, but their predictive power has not increased over time, which is in stark contrast with what we find when partisan labels are incorporated into prediction algorithms. In these richer models with party IDs as one of the predictors, voting behavior is increasingly predictable over time, confirming the well-documented increase in the importance of partisanship in the United States (Bartels, 2000; Levendusky, 2009).

Before describing the results in more detail, it is important to clarify two ideas which have been referred to as 'sorting.' *Ideological sorting* is the notion that both symbolic ideology (i.e., self-identified liberal vs. conservative placement) and operational ideology (i.e., concrete issue positions) are increasingly correlated with partisanship (Fiorina et al., 2011; Hetherington, 2009; Levendusky, 2009; Weber & Klar, 2019). *Social sorting*, on the other hand, is defined as a convergence of social identities and partisan identities, such as race, sexual orientation, religion, occupations, social movements, and so on (Mason, 2016, 2018a; Mason & Wronski, 2018; Weber & Klar, 2019). Both phenomena are believed to entrench partisanship, decreasing the scope for persuasion.

There is little doubt that the ideological distance between Democrats and Republicans has grown (Webster & Abramowitz, 2017) and intra-party heterogeneity has decreased. Supposed consequences include greater partisan animosity (Abramowitz & Webster, 2016; Bougher, 2017; Christenson & Weisberg, 2019; Iyengar et al., 2019), an increase in straight-ticket voting in recent decades (Burden & Kimball, 2009; Jacobson, 2017), and an unprecedented partisan gap in presidential approval ratings (Jacobson, 2019).

But what are the political consequences of social sorting? A natural corollary of social sorting is that demographics should be increasingly prognostic of partisan affiliation. In addition, together with increased party-line voting (Jacobson, 2013), demographic characteristics should also better predict vote choice. We consider a narrow view of social sorting—namely that *demographic group labels* such as age

and race are linked with political behavior—and test whether demographic sorting in vote choice and in partisan identification is growing. Specifically, are demographic characteristics and political opinions/partisanship tied in a way so that belonging to a particular group predicts important political decisions with increasing accuracy?

To our knowledge, existing studies have not quantified the level of demographic labels' joint predictive power over political behavior or analyzed whether it has increased over time. In this paper, we first test whether demographic markers are increasingly informative of vote choice or party ID. For this, we use random forests on data from the ANES (1952–2020) as well as logistic regressions for comparison.

Our contributions in this paper are as follows. First, we show that the classification accuracy under the training/testing paradigm is a quantity of interest that can help validate whether sorting occurs. Contrary to the prevailing narrative, we show that demographics have not become more prognostic of vote choice or partisan identification over time, and standard observable characteristics are of limited usefulness when predicting both political outcomes.

Second, we ask whether other data about voters increase the predictive power of voting models.<sup>1</sup> We systematically investigate how accurately vote choice can be inferred from various combinations of voters' characteristics. We confirm that (1) the predictive power of partisanship is increasing, consistent with the literature, and (2) while other variables additionally contribute to improved predictions, their added value when partisanship is already accounted for is small.

Third, we show that regression models which seem to suggest substantive marginal effects of demographic variables (including race, gender, and education) mask that the overall predictive power of demographics is modest. For example, any model with basic demographics will, as most have rightfully suspected, show that white men with a high school degree are more likely to identify as Republican and to vote Republican compared to women of color with a college degree. However, the best of these modeling efforts will still misclassify 34% of voters in 2020, as we will show in the results section.

## Literature and Hypotheses

Demographic attributes can naturally translate into social identities, and, according to the Columbia School of political behavior, they are reinforced by social networks (Lazarsfeld et al., 1944). Membership in social groups is then said to explain politically relevant beliefs and behaviors, and voting can be viewed as akin to cultural activities which “have their origin in ethnic, sectional, class, and family traditions” (Berelson et al., 1954). In this framework, conformity is the norm.

Some existing research has indirectly hinted that demographic characteristics are only weak predictors of political preferences (Hillygus & Shields, 2014). However, there are several reasons to expect increased predictability from demographics

---

<sup>1</sup> If swing voters are rare (Panagopoulos, 2020), we would expect that vote choices should be easy to predict when enough data about respondents is known.

alone. First, there have been long-term demographic trends such as rising education levels suspected to increase the Democratic foothold (Abramowitz, 2010). Second, political elites have paved the way for a stronger group-party alignment in recent decades, making party positions clearer and more distinguishable to voters. Third, cross-cutting ties between groups have been decreasing (Mason, 2018b), increasing polarization defined in terms of ideological consistency (Lelkes, 2016). Voters are thus expected to vote in line with their perceived group interests and against members of a disliked out-group.

It then seems natural to state that group membership should be increasingly informative of vote choice and party ID. Essentially, this is the wide-spread assumption that “the link between demographic traits and political orientation is so strong that increases in the share of voters from demographic groups associated with support for the Democrats produce proportionate increases in Democratic support,” (a view summarized but not endorsed by Shaw & Petrocik, 2020). Is this assumption defensible? After visiting the existing literature on group-based voting, we formulate three hypotheses to test the implications of the sorting thesis.

## Group-Based Voting

Although social identities that align with political ideology encompass many categories such as religion and interest groups (Abramowitz & Saunders, 2008; Levendusky, 2009), our focus is on the following five demographic markers: race, education, income, age, and gender.<sup>2</sup>

### Race

Although the Democratic Party has had a stable advantage among Black voters for several decades now, there are reasons to expect that the signal from a voter’s race is larger than in the past. First, Trump’s victory in 2016 was a continuation of “the decades-long expansion of Republican support among white working-class Americans” (Carnes & Lupu, 2020). In addition, the GOP is believed to have activated white identity (Abrajano & Hajnal, 2017; Jardina, 2019; Sides et al., 2017, 2019; Tesler, 2016) in response to Barack Obama’s electoral wins. In the case of the Latino voters, the Democratic support varies by ethnicity and generation (Abrajano and Alvarez, 2012), while Asian Americans are, overall, not well courted by either party (Wong et al., 2011). But racial minorities, on average, may have a weaker incentive to vote Republican. Note also that when the Tea Party emerged in 2010, 80–90 percent of its supporters were white (Williamson et al., 2011), strengthening the hypothesis that prediction accuracy based on race could be increasing.

---

<sup>2</sup> While these variables can be viewed as relatively “objective” labels, the strength of individuals’ identification with a specific group is bound to vary.

## Education

Three decades' worth of public opinion demonstrates that those with higher educational attainment increasingly associate with the Democratic party (Pew Research Center, 2018). Conversely, Republicans have been gaining support among those citizens who do not have a college degree in the last decade. The partisan education gap grew during the Obama administration, but note that the relationship between education and voting is sensitive to the inclusion of other variables in a model (Schaffner et al., 2018).

## Income

Income at the individual level predicts vote choice, but there is some disagreement whether class-based voting has been stable (Gelman et al., 2010) or increasing over time (Stonecash, 2000; Bartels, 2006; McCarty et al., 2008).<sup>3</sup> Preferences and voting are typically aligned with people's economic self-interest—for example, Ansolabehere et al. (2006) document that “the difference in the rate of Republican voting between an economic Conservative and an economic Liberal is 31 percentage points.” However, the extent of the importance of economic issues for voting continues to be debated.<sup>4</sup> In the 2016 presidential election, the income effect is believed to have interacted with education. Carnes and Lupu (2020), for example, show that the diploma divide in 2016 “was driven largely by more affluent Americans.”

## Age

Young people tend to lean liberal and support Democratic or progressive candidates. In the 2016 presidential popular vote, the vote margin of Democratic minus Republican votes was 24 for Millennials and 28 for Generation Z (Griffin et al., 2020). Higher age, conversely, is correlated with conservatism and voting Republican,<sup>5</sup> and it has been argued that “the generational divide in American politics today is unprecedented” (Fisher, 2020).

Consider also the self-reported importance of age for perceived common interests. When respondents were allowed to select up to two groups with which they have the most common interests and concerns in a November 2020 YouGov poll, the most frequently mentioned category was “people in the same age group as you,” followed by “people in the same political party.”<sup>6</sup>

<sup>3</sup> The best-known argument comes from McCarty et al. (2008) who argue that there has been growing “stratification of partisanship by income,” with high-income voters increasingly voting Republican.

<sup>4</sup> The relationship between income and Republican partisanship at the individual level, while robust nationally, is moderated by local context according to an analysis of voter files (Hersh and Nall, 2016).

<sup>5</sup> Williamson et al. (2011) found that at least 75% of Tea Party supporters were over 45 years old.

<sup>6</sup> Respondents were asked directly: “Would you say that you share a lot of common interests and concerns with other people of people who are [SAME GROUP], or would you say that age is not really relevant?”. In this context, respondents suggested that class (“people who have about the same amount of money as you”), ethnicity, and geographic proximity were less indicative of common interests than age (YouGov, 2020).

## Gender

Voting patterns in exit polls suggest that men are more likely to vote Republican, but in models that control for sexist attitudes, gender does not appear to predict vote choice (Bracic et al., 2019). At the same time, gender interacts with race. For example, Junn (2017) reports that white women voted Republican in 2016 in line with their behavior in prior elections.

While we believe that these are among the most fundamental demographic characteristics, we also estimate a model with an extended definition of demographics, using whether the respondent lives in an urban/suburban/rural area, census region, political south/nonsouth, working status, religion, church attendance, home ownership, and marital status.<sup>7</sup>

## Hypotheses

Our objective is to test whether demographic sorting has observable implications for partisanship and voting. We argue that the testable implication of demographic sorting is an increase in the predictive power of demographics on both variables. Based on the summarized relationships between demographics and political behavior, we derive the following hypothesis:

**Hypothesis 1A (Rising Vote Choice Predictability)** Vote choice will become increasingly predictable based on voters' demographic features alone (with other information about voters withheld).

**Hypothesis 1B (Rising Party ID Predictability)** Partisanship will be easier to infer based on voters' demographics relative to earlier decades.

Beyond demographics, a long scholarly tradition leads us to expect that an explicit self-reported party label should be a powerful signal of vote choice. Campbell et al. (1960) described the psychological, unthinking allegiance to parties with the following consumer analogy: “[I]ike the automobile buyer who knows nothing of cars except that he prefers a given make, the voter who knows simply that he is a Republican or Democrat responds directly to his stable allegiance” (p. 136). On the basis of this argument, the evidence that voter loyalty is increasing (Burden & Kimball, 2009; Jacobson, 2017), and similar findings (Bartels, 2000), we propose the next hypothesis:

**Hypothesis 2 (Increasing Importance of Party ID)** Including explicit party ID will dramatically raise accuracy relative to sparser models using only demographics, and predicting voting decisions with party ID will grow increasingly easy over time.

---

<sup>7</sup> See the Supporting Information for full details.

Demographics Only		Demo. + PID	Demo. + PID + Issues	All Covariates
Demographics		Demographics	Demographics	Demographics
Basic	+ Extension	7-pt Party ID	7-pt Party ID	7-pt Party ID
Age Gender Race Income Education	Urban Census region Political south Work status Religion Church attd. Home own. Marital status		Issue Positions	Issue Positions
				Everything Else

Fig. 1 Four nested specifications and corresponding variables, visualized

Finally, given that parties are seen as ideological brands (Egan, 2013; Woon & Pope, 2008) and voters now generally belong to the “correct” party (Levendusky, 2009) we propose the final hypothesis:

**Hypothesis 3 (Sufficiency of Party ID)** Beyond the initial sets of features (party ID and demographics), other voter characteristics, such as issue positions, will contain non-zero but substantively small diagnostic information about vote choice.

### Data and Methodology

We use the American National Election Studies (1952–2020), and our target variable for prediction is (1) presidential vote choice, which is self-reported, subsetted to respondents who voted for either a Democratic or a Republican candidate, and (2) binary party ID, also self-reported, of either being a Democrat or a Republican.<sup>8</sup>

For training models in the training/testing paradigm, we use (1) random forests, a commonly used method for aggregating predictions from regression and classification trees (Breiman, 2001; Montgomery & Olivella, 2018), and (2) logistic regressions, as both outcome variables are binary. In random forests, individual trees are estimated by sequentially splitting the data on the basis of an optimally chosen cut-off point of the most informative variable. To remove excessive dependence of tree structures on the algorithmic decisions early in the splitting process, a subset of observations and predictors is drawn each time a new classification tree is estimated. An aggregation of trees corrected for inter-tree correlation is the random forests (RF). Whether it be random forests or logistic regressions, models are trained and fitted separately for each survey wave.

<sup>8</sup> Note that when the outcome variable is presidential vote choice, we use all presidential election years. For binary party ID, we use all available ANES years. Because ANES only conducted small-scale pilots in mid-term years after 2004, years such as 2006, 2010, 2014, and 2018 are missing from the results.

To further investigate the extent to which vote choice is inferrable based on voters' observable characteristics, we use four nested variable specifications for the analysis. Figure 1 shows the labels of each specification and the variables included. Given that the survey questionnaires reflect the issue cleavages of the day, such as the Iraq war or the Affordable Care Act, the third and the fourth specifications will usually consist of imperfectly overlapping sets of variables for each wave. We include these specifications for benchmark purposes.

All categorical variables are converted into dummy variables, including a variable to represent nonresponse missing values. Variables with near-zero variance at the 1% level or variables with more than twenty different responses, such as ZIP codes, are dropped to guard against too much sparsity. Only clearly continuous variables—such as age, number of children, or amount donated to political campaigns—are kept as continuous.<sup>9</sup> Nonresponses are treated as a separate category instead of listwise deletion if the variable is categorical.

After cleaning, the data is split into training and testing datasets with an 80:20 ratio. Using the *caret* and the *ranger* package in R, we run an election-by-election class prediction via (1) random forests (Breiman, 2001; Kuhn, 2008; Wright & Ziegler, 2017) and (2) logistic regressions. All code is publicly available at a GitHub repository: <https://github.com/sysilviakim/surveyML>.

Random forests are tuned before a final (best-fitting) model is chosen out of a set of candidate models. The best model is selected on the basis of its performance in the training set while two hyperparameters that may impact model performance are tuned via tenfold cross-validation.<sup>10</sup>

## Results

### How Much Can Demographics Alone Predict Vote Choice and Partisan Affiliations?

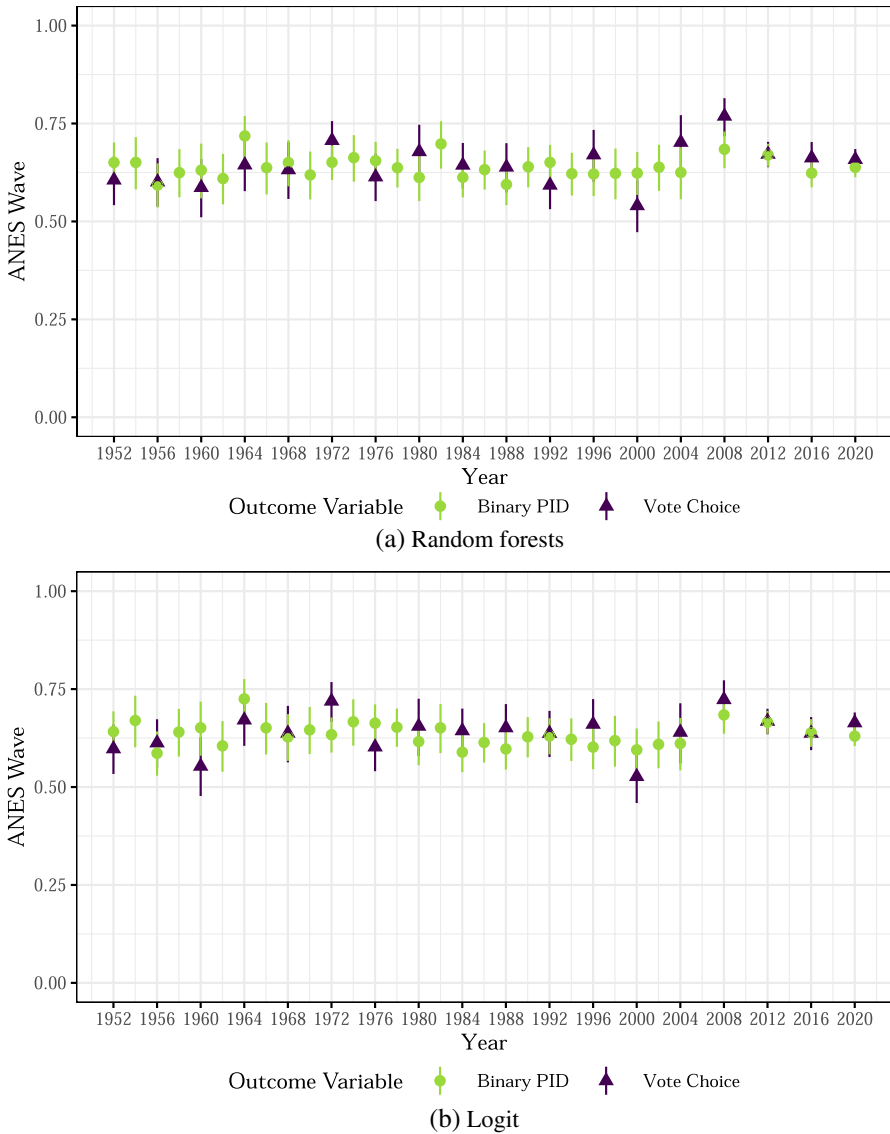
Before we evaluate the evidence for our hypotheses about over-time dynamics, we first quantify how well, on average, demographics can predict political behavior. We find that when using just demographics, the accuracy for vote choice predictions is generally low, typically less than 65%. The numbers are very similar when the target prediction is binary party IDs.

Figure 2 shows the time-series plots of out-of-sample accuracies over time, with the top panel showing random-forest based accuracy rates and the bottom panel showing logistic regression based accuracy rates over time. For both panels, models

<sup>9</sup> We chose to encode ordered categorical variables (such as education and party ID) as sets of binary features for several reasons. First, this allows us to be consistent over various years. Second, we avoid assuming that items measured on an ordered scale have linear or additive effects. Finally, if ordered categorical variables are treated as continuous, item nonresponses would be dropped, but one-hot encoding allows us to keep all survey respondents even when their responses contain missing values, because a separate binary variable is created (e.g., for respondents with a missing income category).

<sup>10</sup> See Neunhoeffer and Sternberg (2019). In our context, it is a procedure for model tuning that uses 80% of the training data.





**Fig. 2** Out-of-sample prediction accuracy of demographics on political behavior, 95% confidence intervals, random forests and logistic regressions

by each year are estimated using only information on respondents' gender, race, education, income, and age.

The average accuracy across all years when predicting vote choice is 63.9% with random forests, and 64.5% when logit models are estimated instead. When predicting binary party ID, the average accuracy for random forests is 63.4%, whereas for

**Table 1** Predictive Accuracy for Vote Choice and Party IDs Using Only Demographics, Regressed Over Year

	Dependent variable: predictive accuracy			
	Vote choice		Party ID	
	Main demo.	Extended	Main demo.	Extended
(a) Random forests				
Year	0.001 (0.001)	0.0004 (0.0005)	-0.0003 (0.0003)	-0.0001 (0.0003)
Observations	18	18	31	31
Adjusted R <sup>2</sup>	0.020	-0.012	-0.001	-0.033
(b) Logistic regressions				
Year	0.001 (0.001)	0.001 (0.001)	-0.000001 (0.0003)	-0.0001 (0.0003)
Observations	18	18	31	31
Adjusted R <sup>2</sup>	0.103	0.165	-0.034	-0.032

\* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$

logistic regressions, it is 63.9%.<sup>11</sup> These numbers will probably strike most readers as low. While better than random guesses,<sup>12</sup> they seem less impressive relative to what could be expected considering the emphasis that the literature places on demographic variables.

Is this because the five demographic traits are too restrictive? As aforementioned and shown in Fig. 1, we also estimate an extended demographic model, where, in addition to the five variables, we add a variety of other demographic variables. The results are not very different. With the extended demographics-only model, the average accuracy for vote choice is 67.4% with random forests and 66.8% with logistic regressions. When predicting binary party ID, the average accuracy for random forests is 67.2% and 67.0% for logistic regressions. Although slightly better than with just five variables, out-sample predictive accuracies are low for both outcome measures and regardless of the algorithm choice.

## No Evidence of Demographic Sorting Over Time

**Hypothesis 1 (Increasing Demographic Sorting)** The next question is then whether, per the first hypothesis, the predictive power of demographics has increased over time. Can demographic labels alone predict presidential vote choice or partisan affiliations better in the polarized era compared to the past?

<sup>11</sup> If we use 3-point party IDs instead of binary PID, so that Independents are also used in a multi-class prediction, the accuracy is 56.2% for random forests and 56.5% for logistic regressions. When 7-point party ID is used instead, the accuracy is 25.2% for random forests and 26.0% for logistic regressions.

<sup>12</sup> For years when election results were not close, 50% is naturally not a fair benchmark for the two party vote. We provide the baseline two-party vote share for each presidential election in Appendix B, which further decreases the value that demographics bring to the prediction (compared to a prediction rule where, for each respondent, we would have naively predicted a vote for the winning candidate).

The time trends shown in both panels of Fig. 2 do not suggest that this is the case. Test-set accuracies from 1952 to 2020 for both outcome variables are remarkably stable. Table 1 tests this as regressions for both outcomes, whether basic or extended set of demographics was used, and for both choices of algorithms. None of the over-time trends are statistically significant. Again, the results are not dependent on the choice of the training model.

The results suggest that the electorate has not become more polarized along demographic lines in a way that is *informative about voting behavior*.<sup>13</sup> Note that this is not about a specific demographic group (e.g., the marginal contribution of white working-class men in their 60 s to a party voting bloc) but the overall joint prediction ability of combinations of demographics, which may explain the disparity between the results of this paper and the literature's recent focus.

### Growing Marginal Effects Do Not Imply Rising Joint Accuracy

Why is the overall accuracy low and stable when other work suggests that demographic sorting is increasing? One reason may be that counteracting forces are playing out simultaneously: indeed, between 2012 and 2020, the electorate seems to have been polarized by education, but it also appears to have racially depolarized.

A set of marginal effects from election-by-election models displayed in Fig. 3 shows that race and gender have for decades been significantly associated with both party ID and with vote choice. Moreover, the influence of education on the predicted probability that a respondent would vote Republican (top panel) or identify as a Republican (bottom panel) has increased in recent years. But these panels mask the lack of an increase in model fit, which we documented in the preceding subsection.

Figure 3 thus conveys that statistical significance does imply that a set of variables is highly diagnostic or particularly useful for better predictions (as has been noted in other contexts; see e.g. Lo et al., 2016). Because plots of marginal effects, or indeed crosstabulations, can sometimes look striking, it seems worth restating that growing marginal effects do not, in the absence of other information, imply that any given model would yield accurate predictions.

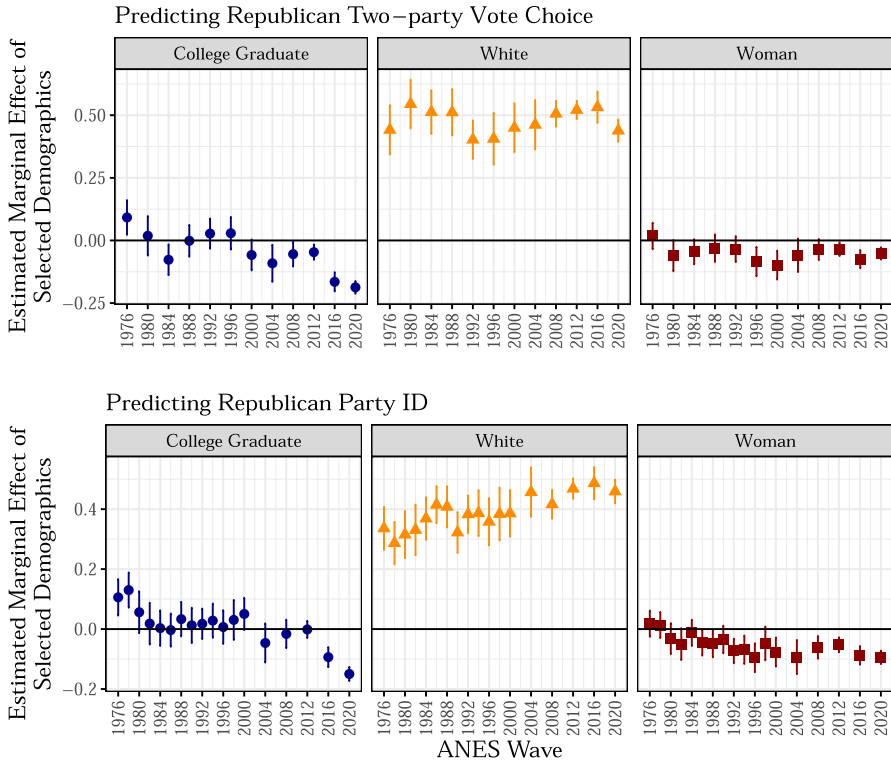
### Prognostic Power of Party ID and Ideology on Vote Choice

**Hypothesis 2 (Increasing Party ID Sorting)** Next, we turn to tests of our second hypothesis which stated that partisan identification would become more informative over time for two-party vote choice, consistent with the existing literature.

Tree-based models that take advantage of data on respondents' explicit partisan identification, measured on a seven-point scale, are presented (with other models

---

<sup>13</sup> Note that AUC (in Appendix), over time, seems to be increasing for vote choice unlike accuracy. The slope is 0.0019 with a standard error of 0.0005 ( $p$ -value < 0.01), hinting that over all possible threshold values, the ability to separate two-party vote choices has increased over time. But when classifying vote choices, researchers rarely use a threshold other than 50%, so we present accuracy as our main results. Similarly, for binary PID, the slope is 0.0019 with a standard error of 0.0004 ( $p$ -value < 0.01).



**Fig. 3** Marginal effects of college graduation status, white, and female on republican vote choice and binary PID, simple OLS with demographics

included for comparison) in Fig. 4 which shows point estimates of three performance metrics (accuracy, AUC, and the F-1 score).<sup>14</sup> These performance metrics are also shown for the remaining nested specifications.

We find that partisanship, jointly with basic demographics, has indeed become a significantly more prognostic variable over time, confirming that our approach yields results consistent with the existing literature. Before the 1992 election, PID-based accuracy never exceeded 85%. However, starting with Bill Clinton’s re-election, the same specification generally classifies at least 90% of voting decisions correctly. We thus find empirical support for our second hypothesis, which aligns with the well-established results of an increasingly polarized U.S. electorate.

Of course, the predictions based on the second specification leave on the table other knowable attributes of voters. We explore the implications of their inclusion, which brings us to our final hypothesis.

<sup>14</sup> Note that the first model, demographics only, uses the basic specification with five variables. Logistic regressions perform poorly with models with a large number of covariates such as the all-covariates model, and thus are not presented here.

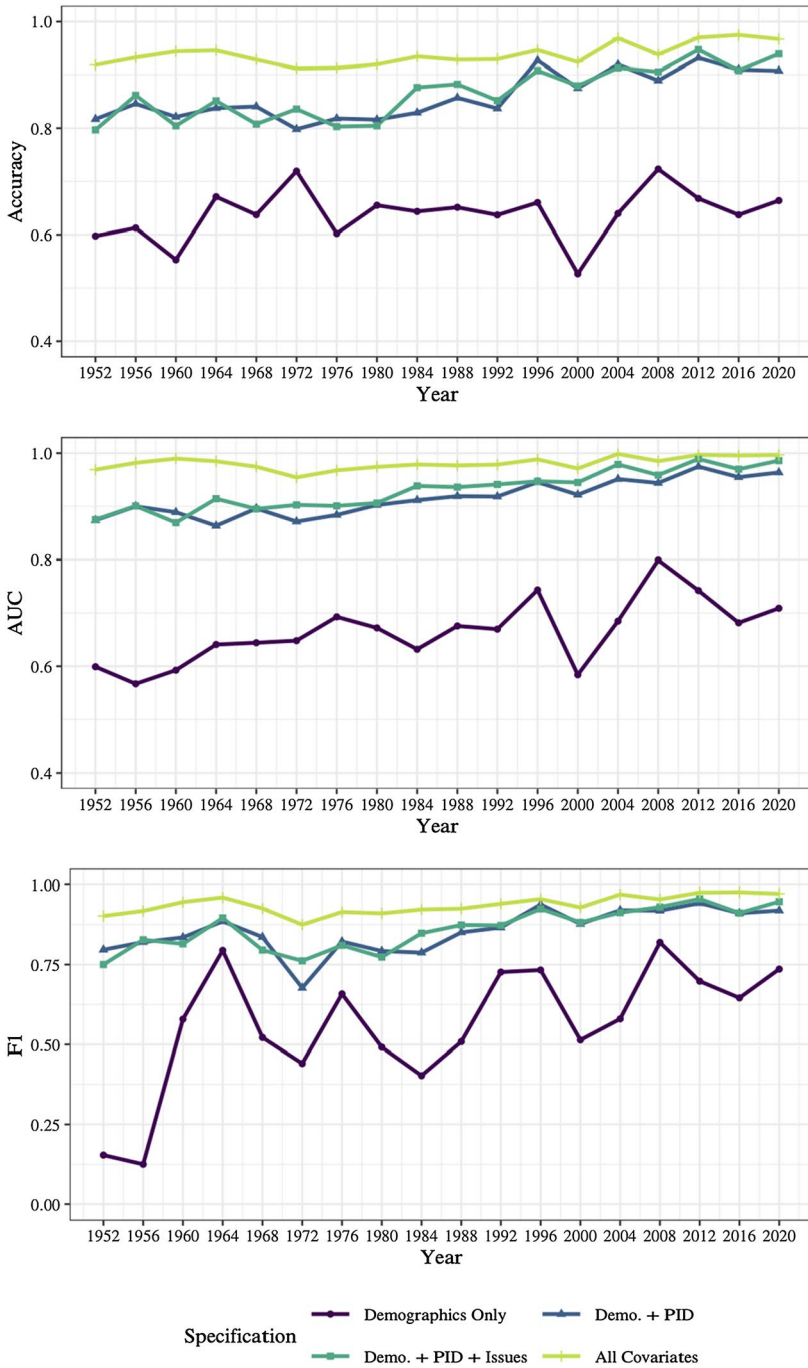


Fig. 4 Performance of presidential vote prediction over time (random forests, accuracy/AUC/F1 scores)

**Hypothesis 3 (Sufficiency of Party ID)** While partisanship is now more prognostic of vote choice, other factors do continue to provide additional useful information for inferring respondents' vote choice, as we can see in Fig. 4.

More specifically, on average, accuracy of ANES predictions improves by 0.5 percentage point when issue variables are added on top of demographics and party ID. In addition, accuracy improves by 7.9 percentage points once all other variables have been added. Some examples of variables included in the fourth and final specification are non-policy opinions. In 2016, for example, a useful feature was the perception about whether the Democratic presidential candidate could be described as honest.

The patterns uncovered by these models suggest that it is possible to glean information about voters' behavior even after accounting for their partisanship. Views on policy issues consistently reveal more information about behavior, above and beyond partisanship. Moreover, other questions asked on public opinion surveys (occupation, subjective class identification, group attitudes, political knowledge, media consumption, beliefs, perceptions, and so on) still contain additional information that can be used to improve predictions about voters' behavior.

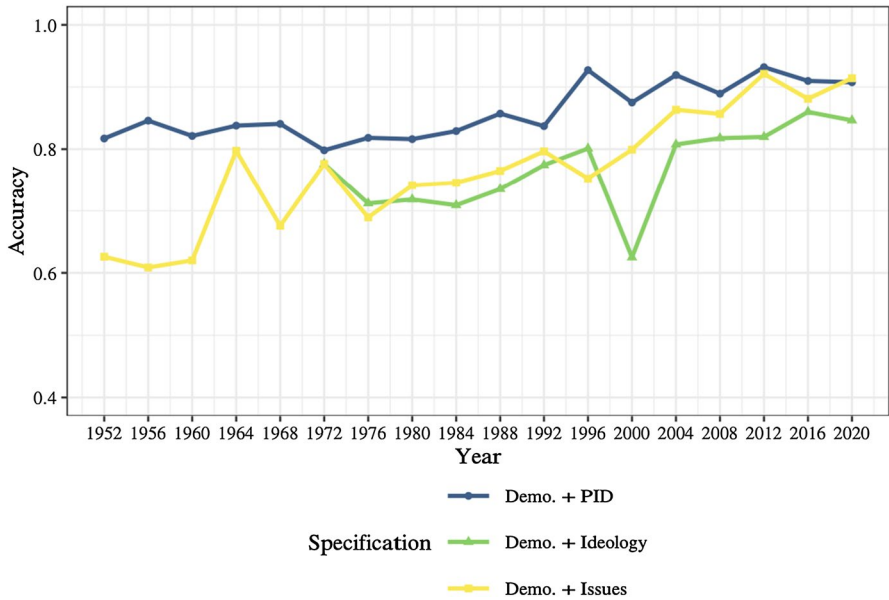
However, it is clear that in recent years, the added value from the set of all variables in terms of prediction is, on average, decreasing. Once party ID is accounted for along with demographics, the ability of other variables to be put to good use in better predicting vote choice is more limited compared to the era with lower mass-level polarization.

### Party ID vs. Symbolic Ideology vs. Operational Ideology

Given the results above, it is also worth digressing to investigate the predictive power of three distinct concepts which are sometimes confused: explicit party ID, self-identified symbolic ideology along the liberal-conservative scale, and operational ideology, or specific positions on issues (Ellis & Stimson, 2012). Figure 5 shows the accuracy of predictions based on the following models: demographics and seven-point party ID (also displayed in Fig. 4), demographics and three-point symbolic ideology (liberal vs. moderate vs. conservative<sup>15</sup>), and demographics and operational ideology, or all issue questions. Note that some caution in interpretation must be forewarned due to the number of varying covariates between this model.

Three patterns are worth noting. First, symbolic ideology is nearly as informative as operational ideology (although its predictive power relative to issues decreased in 2020). Second, in spite of a large amount of information available in the data in the operational ideology model, issues typically underperform compared to party ID. Finally, the convergence of the predictive power based on demographics and operational ideology vis-à-vis the demographics and party ID model highlights the growing alignment between partisanship and policy views.

<sup>15</sup> Models begin in 1972 when ANES starts providing the ideological self-identification question.



**Fig. 5** Performance of presidential vote prediction over time. Comparison of party ID, symbolic ideology, and operational ideology. Accuracy is the (random forest based) percent of correctly predicted observations

This exercise makes it clear that party labels which voters give themselves trump both types of ideologies. In the next section, we turn back to demographics and describe how their relative importance has changed over time.

### Discussion: A Contested Role of Demographics

The degree of group-level voter loyalty has substantial implications for representation. Per the typical rational choice model, campaigns will focus on catering to persuadable voters rather than pandering to a part of the electorate that will, conditional on turning out, behave predictably. If a voting bloc is “too reliable,” a normative concern is that the group’s interests will not be represented adequately relative to its size and importance. For example, the Democratic party has been criticized for not giving priority to issues that matter to Black voters, a key voting bloc that delivered the Biden victory (Scott, 2020).

Turning the clock back a few years, Donald Trump’s unexpected victory in 2016 left the media and researchers searching for explanations of the winner’s appeal, and they often zoomed in on the voting blocs that supported Trump, such as white

voters without a college degree.<sup>16</sup> A common concern has been that existing political cleavages between demographic groups are widening, and many of the post-election explanations focused on the deepening partisan divide by demographics such as race or education (Morgan & Lee, 2018; Porter, 2016).

Along these lines, Sides (2017) writes that “[t]he Democratic Party has an increasing advantage among nonwhite people. Among Hispanics, Democrats outnumbered Republicans by 23 points in 2002 but 36 points in 2016.” The reduction of the white population and increasing racial diversity of the electorate (Teixeira et al., 2015) have been used to project the advent of the Democratic Party.

But other scholars have cautioned that belonging to a social group “does not necessarily prescribe a specific political outlook” (Huddy, 2018). Clearly, unexpected behaviors from some groups can lead to wrong inferences about election competitiveness; for instance,

McCall and Orloff (2017) have observed that:

... some Democratic commentators bemoaned the fact that a majority of white women had voted for Trump, and called it a kind of betrayal, underlining their expectation that women would naturally, on the basis of their gender interests and identity, support a woman with politics and policies understood to be women-friendly.

This is particularly important because, in this paper, we are using demographic *labels*, which are self-reported categorizations and not the *degrees* to which individuals identify with the group labels. To put it another way, objective inclusion in groups is not equal to the internalized sense of membership (Huddy, 2013), so demographic labels are not equal to the social identity and affinity discussed in Green et al. (2002), Achen and Bartels (2016), or Mason (2015, 2016). Therefore, theoretically, our results show that the demographic sorting based on observable ‘labels’ does not extend to a better prediction of vote choices or partisan affiliations.

Our results also have substantial implications for practitioners and the public. The results could help dispel the myth that demographics are deterministic indicators of vote choices. Such myths are perpetuated due to the horse-race coverage of pre-election and postelection polls based on group membership. They have the potential to wreak real-world damage by widening the emotional gap between groups by wrong ecological inferences.

All this is not to say that demographics should be disregarded altogether when campaign strategies are formulated. Because we investigate vote choice, the data sample itself is conditional on turnout, one of the key factors that determine a group’s contribution to the voting bloc (Axelrod, 1972). Therefore, demographics can still play an important role by influencing turnout. For example, Krupnikov and Piston (2015) show that when there are Black candidates on the ballot, racial prejudice may prompt even co-partisan voters to stay home instead of turning out to vote.

---

<sup>16</sup> Abramowitz and McCoy (2019) conclude that Trump’s 2016 campaign slogan of “Make America Great Again” has successfully pulled the white working-class, especially those without a college degree, away from the Democratic party. See also Lamont et al. (2017) and McQuarrie (2017).



This indicates that if demographics are strong determinants of turnout in certain races, the attempt to see the relationship between demographics and vote choice may suffer from bias induced by conditioning on a collider. We leave further probes for future research.

## Conclusion

Demographic attributes can function as markers of social identities, and membership in these social groups does, to an extent, carry political meaning. Voters' demographic characteristics help improve vote choice predictions compared to random guesses—given five key demographic characteristics, the probability of an accurate prediction with a random forests model is on average 63.9% for vote choice and 63.4% for partisan affiliations. This, although higher than 50%, suggests that for most people, memberships in their income group, age group, gender, education group, or even ethnic group are not electorally 'sorted' strongly enough to translate to particularly accurate signals about their voting decisions, or indeed partisan identification. The results are robust to expanding the set of demographic variables used for prediction.

Moreover, the accuracy of predicted outcomes—party ID or vote choice—inferred based on voters' demographic attributes has not grown over the years. These results hold whether we use tree-based models or logistic regressions in training our models. Our results are therefore not consistent with the first hypothesis: demographic sorting is neither electorally consequential nor increasing.

We validate our first result by showing results from four nested model specifications, increasingly incorporating more information about respondents. In line with existing findings showing higher polarization, we show that once partisanship is no longer withheld from the set of predictors, we do observe, as expected, a massive increase in accuracy. Partisanship's diagnosticity in absolute terms is unmatched by either symbolic or operational ideology, although issue positions and partisanship's predictive power have converged in recent years. Furthermore, inferring vote choice with just the combination of demographics and party ID becomes easier over time (with a plateau occurring in the late 1990s).

We also note that predictions based on partisanship and demographics can further be improved by incorporating data on respondents' issue positions. Adding extra variables (including non-policy features such as voters' political knowledge, attitudes tapping into identity considerations, and other survey instruments) to the set of model features also generally yields higher accuracy. However, once partisanship has been accounted for, the added value from additional variables is decreasing over time. In recent years, there has been only minimal gain from richer specifications once partisanship and demographics are accounted for.

Considering all this, we conclude that demographic reconfigurations of partisan coalitions have not translated into more predictable voting behavior. Without information about respondents' partisanship, even sophisticated random forests models typically only achieve out-of-sample accuracy up to 65%, and this performance metric has not increased over the last seventy years. As the predictive power of party ID

grows stronger, it dominates the signal from other covariates, diminishing their marginal predictive power. Therefore, while our results validate scholarly findings on ideological sorting and polarization, we find no support for vote-based demographic sorting.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11109-022-09816-z>.

**Acknowledgements** We thank Jeff Gill, Shiro Kuriwaki, R. Michael Alvarez, Patrick Egan, Matt Grossman, Jake Haselswerdt, Jennifer Wolak, Hye Young You, and the participants at MPSA 2021, PolMeth Europe 2021, the NYU American Politics workshop, and the American University works-in-progress series for comments. This paper was previously circulated under the working title, *The Divided (But Not More Predictable) Electorate: A Machine Learning Analysis of Voting in American Presidential Elections*.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abrajano, M., & Hajnal, Z. L. (2017). *White backlash: Immigration, race, and American politics*. Princeton University Press.
- Abrajano, M., & Michael Alvarez, R. (2012). *New faces: The Hispanic Electorate in America*. Princeton University Press.
- Abramowitz, A. (2010). *The disappearing center: Engaged citizens, polarization, and American democracy*. Yale University Press.
- Abramowitz, A., & McCoy, J. (2019). United States: Racial resentment, negative partisanship, and polarization in Trump's America. *The ANNALS of the American Academy of Political and Social Science*, 681(1), 137–156.
- Abramowitz, A. I., & Saunders, K. L. (2008). Is polarization a myth? *The Journal of Politics*, 70(2), 542–555.
- Abramowitz, A. I., & Webster, S. (2016). The rise of negative partisanship and the nationalization of U.S. elections in the 21st century. *Electoral Studies*, 41, 12–22.
- Achen, C., & Bartels, L. (2016). *Democracy for realists*. Princeton University Press.
- Ansolabehere, S., Rodden, J., & Snyder, J. M. (2006). Purple America. *Journal of Economic Perspectives*, 20(2), 97–118.
- Axelrod, R. (1972). Where the votes come from: An analysis of electoral coalitions, 1952–1968. *The American Political Science Review*, 66(1), 11–20.
- Bartels, L. M. (2000). Partisanship and voting behavior, 1952–1996. *American Journal of Political Science*, 44(1), 35–50.
- Bartels, L. M. (2006). What's the matter with what's the matter with Kansas? *Quarterly Journal of Political Science*, 1(2), 201–226.
- Berelson, B. R., Lazarsfeld, P. F., & McPhee, W. N. (1954). *Voting: A study of opinion formation in a presidential campaign*. University of Chicago Press.

- Bougher, L. D. (2017). The correlates of discord: Identity, issue alignment, and political hostility in polarized America. *Political Behavior*, 39(3), 731–762.
- Bracic, A., Israel-Trummel, M., & Shortle, A. F. (2019). Is sexism for white people? Gender stereotypes, race, and the 2016 presidential election. *Political Behavior*, 41(2), 281–307.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Burden, B. C., & Kimball, D. C. (2009). *Why Americans split their tickets: Campaigns, competition, and divided government*. University of Michigan Press.
- Campbell, A., Converse, P. E., Miller, W. E., & Stokes, D. E. (1960). *The American Voter*. NY: John Wiley.
- Carnes, N., & Lupu, N. (2020). The white working class and the 2016 election. *Perspectives on Politics*, 19, 55–72.
- Christenson, D. P., & Weisberg, H. F. (2019). Bad characters or just more polarization? The rise of extremely negative feelings for presidential candidates. *Electoral Studies*, 61, 102032.
- Egan, P. J. (2013). *Partisan priorities: How issue ownership drives and distorts American politics*. Cambridge University Press.
- Ellis, C., & Stimson, J. A. (2012). *Ideology in America*. Cambridge University Press.
- Fenno, R. F. (1978). *Home style: House members in their districts*. Longman.
- Fiorina, M. P., Abrams, S. J., & Pope, J. (2011). *Culture war?* Longman.
- Fisher, P. (2020). Generational cycles in American politics, 1952–2016. *Society*, 57, 22–29.
- Gelman, A., Kenworthy, L., & Yu-Sung, Su. (2010). Income inequality and partisan voting in the United States. *Social Science Quarterly*, 91(5), 1203–1219.
- Green, D., Palmquist, B., & Schickler, E. (2002). *Partisan hearts and minds*. New Haven, CT: Yale University Press.
- Griffin, R., Frey, W. H., & Teixeira, R. (2020). America's electoral future: The coming generational transformation. Center for American Progress. Retrieved from <https://www.brookings.edu/research/americas-electoral-future-the-coming-generational-transformation/>.
- Hersh, E. D. (2015). *Hacking the electorate: How campaigns perceive voters*. Cambridge University Press.
- Hersh, E. D., & Nall, C. (2016). The primacy of race in the geography of income-based voting: New evidence from public voting records. *American Journal of Political Science*, 60(2), 289–303.
- Hetherington, M. J. (2009). Putting polarization in perspective. *British Journal of Political Science*, 39(2), 413–448.
- Hillygus, D. S., & Shields, T. G. (2014). *The persuadable voter: Wedge issues in presidential campaigns*. Princeton University Press.
- Huddy, L. (2013). From group identity to political cohesion and commitment. In L. Huddy, D. O. Sears, & J. S. Levy (Eds.), *The Oxford handbook of political psychology* (2nd ed.). Oxford University Press.
- Huddy, L. (2018). The group foundations of democratic political behavior. *Critical Review*, 2(30), 71–86.
- Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22(1), 129–146.
- Jacobson, G. C. (2013). Partisan polarization in American politics: A background paper. *Presidential Studies Quarterly*, 43(4), 688–708.
- Jacobson, G. C. (2017). The triumph of polarized partisanship in 2016: Donald Trump's improbable victory. *Political Science Quarterly*, 132(1), 9–41.
- Jacobson, G. C. (2019). Extreme referendum: Donald Trump and the 2018 midterm elections. *Political Science Quarterly*, 134(1), 9–38.
- Jardina, A. (2019). *White identity politics*. Cambridge University Press.
- Junn, J. (2017). The Trump majority: White womanhood and the making of female voters in the US. *Politics, Groups, and Identities*, 5(2), 343–352.
- Krupnikov, Y., & Piston, S. (2015). Racial prejudice, partisanship, and white turnout in elections with black candidates. *Political Behavior*, 37(2), 397–418.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*. <https://doi.org/10.18637/jss.v028.i05>
- Lamont, M., Park, B. Y., & Ayala-Hurtado, E. (2017). Trump's electoral speeches and his appeal to the American white working class. *The British Journal of Sociology*, 68(S1), S154–S180.

- Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). *The people's choice: How the voter makes up his mind in a presidential campaign*. Sloan and Pearce: Duell.
- Leikes, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*, 80(S1), 392–410.
- Levendusky, M. (2009). *The partisan sort: How liberals became democrats and conservatives became republicans*. University of Chicago Press.
- Levendusky, M. S. (2018). Americans, not partisans: Can priming American national identity reduce affective polarization? *The Journal of Politics*, 80(1), 59–70.
- Lo, A., Chernoff, H., Zheng, T., & Lo, S.-H. (2016). Framework for making better predictions by directly estimating variables' predictivity. *Proceedings of the National Academy of Sciences*, 113(50), 14277–14282.
- Mason, L. (2015). "I Disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, 59(1), 128–145.
- Mason, L. (2016). A cross-cutting calm: How social sorting drives affective polarization. *Public Opinion Quarterly*, 80(S1), 351–377.
- Mason, L. (2018a). Losing common ground: Social sorting and polarization. *The Forum*, 16(1), 47–66.
- Mason, L. (2018b). *Uncivil agreement: How politics became our identity*. University of Chicago Press.
- Mason, L., & Wronski, J. (2018). One tribe to bind them all: How our social group attachments strengthen partisanship. *Political Psychology*, 39(S1), 257–277.
- McCall, L., & Orloff, A. S. (2017). The multidimensional politics of inequality: Taking stock of identity politics in the US Presidential election of 2016. *The British Journal of Sociology*, 68(S1), S34–S56.
- McCarty, N., Poole, K. T., & Rosenthal, H. (2008). *Polarized America: The dance of ideology and unequal riches*. MIT Press.
- McQuarrie, M. (2017). The revolt of the Rust Belt: Place and politics in the age of anger. *The British Journal of Sociology*, 68(S1), S120–S152.
- Montgomery, J. M., & Olivella, S. (2018). Tree-based models for political science data. *American Journal of Political Science*, 62(3), 729–744.
- Morgan, S., & Lee, J. (2018). Trump voters and the white working class. *Sociological Science*, 5, 234–245.
- Neunhoeffer, M., & Sternberg, S. (2019). How cross-validation can go wrong and what to do about it. *Political Analysis*, 27(1), 101–106.
- Panagopoulos, C. (2015). All about that base: Changing campaign strategies in US Presidential elections. *Party Politics*, 22(2), 179–190.
- Panagopoulos, C. (2017). *Political campaigns: Concepts, context, and consequences*. Oxford University Press.
- Panagopoulos, C. (2020). *Bases loaded: How US presidential campaigns are changing and why it matters*. Oxford University Press.
- Pew Research Center. (2018). Wide gender gap, growing educational divide in voters' party identification.
- Porter, E. (2016, Dec 13). Where were Trump's votes? Where the jobs weren't. *New York Times*.
- Schaffner, B. F., MacWilliams, M., & Nteta, T. (2018). Understanding white polarization in the 2016 vote for president: The sobering role of racism and sexism. *Political Science Quarterly*, 133(1), 9–34.
- Scott, Eugene (2020). Black voters delivered Democrats the presidency. Now they are caught in the middle of its internal battle. *Washington Post*.
- Shaw, D., & Petrocik, J. (2020). *The turnout myth*. Oxford University Press.
- Sides, John (2017). Race, religion, and immigration in 2016: How the debate over American identity shaped the election and what it means for a trump presidency. *Democracy Fund Voter Study Group*. Retrieved from <https://www.voterstudygroup.org/publication/race-religion-immigration-2016>.
- Sides, J., Tesler, M., & Vavreck, L. (2017). How Trump lost and won. *Journal of Democracy*, 28(2), 34–44.
- Sides, J., Tesler, M., & Vavreck, L. (2019). *Identity crisis: The 2016 presidential campaign and the battle for the meaning of America*. Princeton University Press.
- Stonecash, J. (2000). *Class and party in American politics*. Routledge.

- Teixeira, R., Frey, W. H., & Griffin, R. (2015). States of change: The demographic evolution of the American electorate, 1974–2060. Technical report, Center for American Progress.
- Tesler, M. (2016). *Post-racial or most-racial? Race and politics in the Obama era*. University of Chicago Press.
- Weber, C., & Klar, S. (2019). Exploring the psychological foundations of ideological and social sorting. *Political Psychology*, 40(S1), 215–243.
- Webster, S. W., & Abramowitz, A. I. (2017). The ideological foundations of affective polarization in the US electorate. *American Politics Research*, 45(4), 621–647.
- Williamson, V., Skocpol, T., & Coggin, J. (2011). The tea party and the remaking of republican conservatism. *Perspectives on Politics*, 9(1), 25–43.
- Wong, J. S., Karthick Ramakrishnan, S., Lee, T., & Junn, J. (2011). *Asian American political participation: emerging constituents and their political identities*. Russell Sage Foundation.
- Woon, J., & Pope, J. C. (2008). Made in congress? testing the electoral implications of party ideological brand names. *Journal of Politics*, 70(3), 823–836.
- Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17.
- YouGov. (2020). HuffPost: Common interests. Retrieved from [https://today.yougov.com/topics/politics/explore/topic/American\\_culture](https://today.yougov.com/topics/politics/explore/topic/American_culture).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.