What Can We Learn About How Political Campaigns Activate Attitudes?

Justin Grimmer* William Marble† Cole Tanigawa-Lau‡

First Draft: September 25, 2021
This Draft: September 29, 2021

Abstract

In the study of elections, there is substantial interest in the ability of campaigns and the media to influence the criteria that voters use to evaluate politicians. Despite this widespread interest in "activation" or "priming," there is a lack of consensus on the exact estimand being studied. Further, the literature has not formally analyzed the conditions under which the activation estimand can be identified. In this paper, we draw on the literature to formalize three distinct conceptions of activation. We use these formalizations to analyze commonly used observational and experimental research designs. A key result is that two ignorability assumptions are necessary to assess whether campaigns or the media caused voters to change the weight they place on an issue in their voting decisions. A weaker formulation of activation, based on prediction, relaxes one of the assumptions, at the cost of reduced theoretical interpretability. Our framework organizes the literature on activation and priming, and may help spur development of new research designs that improve estimation of activation effects.

*Professor, Department of Political Science, Stanford University and Senior Fellow at the Hoover Institution. jgrimmer@stanford.edu.

†Postdoctoral Fellow, Center for the Study of Democratic Politics, Princeton University. wmarble@princeton.edu.

‡Ph.D. Student, Department of Political Science, Stanford University. coletl@stanford.edu.

Contents

1 Introduction

A key question in the study of elections is the extent to which campaigns, the media, and other political actors can influence voters' decision-making process. Under the "priming" or "activation" hypothesis, candidates can make implicit or explicit appeals that change the criteria by which voters judge politicians. Such appeals "activate" citizens' attitudes on particular issues and cause voters to bring their candidate evaluations in line with these pre-existing attitudes (Hutchings and Jardina, 2009). This framework has been used to study the influence of news media on public opinion (Iyengar and Kinder, 1987), the effect of racial appeals on candidate choice and attitudes on ostensibly non-racial issues (Mendelberg, 2001; Valentino, Hutchings and White, 2002; Tesler, 2012), the ability of the public to hold politicians accountable (Lenz, 2013), and the causes of particular election outcomes (Sides, Tesler and Vavreck, 2019; Hopkins, 2019).

Despite this large literature, there is little formal discussion of what priming or activation entails, much less the assumptions under which they can be identified using empirical data. Researchers variously refer to priming or activation as changes to the weight that voters place on certain issues, the predictive power of an attitude on vote choice, the correlation between an attitude and vote choice, or the accessibility of certain attitudes for opinion formation.[1] These different definitions correspond to varying levels of specificity regarding mechanisms and suggest alternative research strategies. Given that they specify different phenomena of interest, they also differ in the extent to which alternative mechanisms represents threats to inference.

In this paper, we provide a framework that consolidates the goals of activation and priming analyses, the assumptions needed to sustain inferences, and proper interpretations of the results. We formalize three different notions of activation that we identify in the literature.

In the first formulation, we define activation as the weight that voters attach to certain

[1]This is to say nothing of the broader use of the term "priming" by social psychologists to describe the influence of subtle environmental cues on decision-making—a phenomenon which is also under-theorized (Molden, 2014). For an application of this meaning of "priming" to political behavior, see Berger, Meredith and Wheeler (2008).

issues or attitudes in their vote choice decision. We formalize this notion in the context of a weighted spatial voting model, whereby voters evaluate candidates according to their agreement across a range of issues. The model provides a simple but precise characterization of different mechanisms relating vote choice to issue attitudes. In the model, activation refers to an increase in the weight attached to one dimension. Two other potential campaign effects identified by Lenz (2009)—"learning" and "opinion change"—are also neatly captured by the model. We show that simple regressions of vote choice on issue attitudes do not capture issue weights, but rather a combination of issue weights and candidate positions.

In the second formulation, we relax the behavioral assumptions underlying the spatial voting model and instead examine an attitude's causal effect on vote choice—regardless of the particular mechanism. In this context, activation is defined as a *causal moderation effect*, whereby a campaign event or other stimulus causes an increased causal effect of an attitude on vote choice. This formulation, which is a generalization of the first, highlights that studies of activation often implicitly rely on two ignorability assumptions. First, in order to estimate an attitude's causal effect on vote choice, researchers must assume ignorability of the attitude (perhaps conditional on covariates). Second, in order to study the effect of the campaign stimulus, researchers must assume ignorability of the campaign event. While the latter assumption can be guaranteed in experimental settings, the first is nearly impossible to ensure. To improve the plausibility of this assumption, researchers often use panel data and relate lagged attitudes to contemporaneous vote choice. We use our framework to study this research design and conclude that it presents a trade-off. The assumption of ignorability of attitudes is made more plausible, but the implied estimand is different from the one researchers typically target.

Finally, the third formulation focuses on the *predictive* relationship between an attitude and vote choice, and the potential for campaign events to change this relationship. This formulation is the weakest theoretically, as it has little to say about the effect an attitude has on candidate evaluations. However, because one can assess the predictive capacity of a variable without ignorability assumptions, it is also the most credibly estimable of the three formulations. Identifying "predictive activation" requires an ignorability assumption on campaign

events only. This formulation also suggests that linear regression should not necessarily be the workhorse empirical approach. Instead, researchers should use more flexible estimators that directly optimize predictive accuracy.

Our goal in providing these formalizations is to provide clarity over the different inferential goals at play in the priming and activation literature, and to highlight strengths and weaknesses of different research designs. To that end, we provide suggestions throughout the paper for empirical approaches that more closely align with the target estimand.

2 Outlining the Goals of Activation and Priming Literature

The study of "activation" goes back to foundational works in political behavior. In *The People's Choice*, Lazarsfeld, Berelson and Gaudet (1948) argue that instead of persuading voters, "political campaigns are important primarily because they *activate* latent predispositions" (74, emphasis in original). Subsequent research furthered the view that political opinions were crystallized early in life, leaving little room for campaigns to persuade voters (Campbell et al., 1960). Indeed, for much of the 20th century, scholars largely accepted the "minimal effects" hypothesis about campaigns. Research on political communication was revitalized in the late 1980s, however, with the advent of laboratory studies and theoretical development of the concepts of priming, framing, and agenda-setting (Iyengar and Kinder, 1987). The concept of priming is similar to the idea of "activation" that the early Columbia researchers posited. Priming research largely shows how campaigns and news media can highlight certain issues, thereby causing citizens to bring their evaluations of candidates in line with their prior attitudes on those issues.[2]

Despite the long history of activation, few studies provide a precise definition of what activation entails. The core goal of this paper is to rigorously define several notions of

[2]Some authors prefer the term "priming" while others prefer the term "activation." We have not come across a clear delineation between them, and have found that they are sometimes used interchangeably. For example, Hopkins (2019, 664) writes that "political rhetoric can make people's pre-existing attitudes toward social groups more central in their support for candidates, an effect known as priming or activation." Similarly, Tesler's (2015) study—entitled "Priming Predispositions and Changing Policy Attitudes"—refers to predispositions being "activated" over the course of a campaign (809).

activation and state the assumptions necessary to identify activation using empirical data. Here, we briefly outline each of our three definitions of activation and highlight prior research that employs these definitions.

2.1 Activation as Issue Weights

First, activation may refer to an increase in the weight that voters give a certain issue when forming their opinions. In this notion of activation or priming, overall candidate evaluations are a product of voters' opinions toward candidates on a series of issue dimensions. When an attitude is activated, the importance of that dimension increases relative to other dimensions.

This notion is implied in many studies of priming, especially. Lenz (2009, 822) summarizes priming research thus:

> Researchers generally test whether an increase in the prominence of an issue leads individuals to increase the weight given to the issue when evaluating rival candidates or incumbent politicians. They measure such increases by regressing presidential approval or vote choice on a series of policy attitudes. The coefficients from these regressions, also called 'issue weights,' are interpreted as reflecting the importance people place on each issue when evaluating the president or deciding for whom to vote. Researchers then examine whether these issue weights vary with the prominence of the issues.

Just as Lenz describes, in their agenda-setting book *News That Matters*, Iyengar and Kinder (1987, 66) refer to priming in terms of the "weight" given to issues highlighted by media: "If the priming hypothesis is correct, we should find that viewers who were shown stories about a particular problem gave more weight to the president's performance *on that problem* when evaluating the president's overall performance" (emphasis in original). Other studies echo the idea that priming involves increasing the weight placed on an issue (Miller and Krosnick, 2000; Hart and Middleton, 2014).

The idea of issue weights inherently involves an underlying behavioral model of voters' decision-making process, whereby voters consider their opinions over a range of issues. These

issues may be directly policy relevant (e.g., How well does this candidate's stance on healthcare align with my own?) or may reflect identity-based concerns (e.g., Does this candidate's rhetoric reflect my idea of who should be included and excluded from the polity?). Regardless of the content of these issues, priming involves changing the importance placed on one of these dimensions relative to another.

In Section 3, we provide one potential formalization of voters' behavior and use that structure to analyze research strategies employed in studies of priming and activation. In the model, citizens' overall evaluations of a politician are a weighted sum of their evaluations of the candidates over a host of issues. Priming or activation occurs when the weight that citizens place on one of these issues increases.[3] A key result from our model is that simple correlations between vote choice and attitudes are insufficient to identify issue weights. The reason is that this correlation depends both on an issue's importance and candidates' positioning on the issue. An increased correlation between issue attitudes and vote choice—which is generally the key statistical test in priming and activation studies (Iyengar and Kinder 1987, 156; Valentino, Hutchings and White 2002, 81; Hart and Middleton 2014, 586)—could reflect either greater weight on the issue or greater perceived distance between the candidates on that issue.

2.2 Activation as a Bundle of Campaign Effects

A second notion of "activation" abstracts away from the rigid behavioral model implicitly invoked in the activation-as-issue-weights definition. Instead, it focuses on a change in an attitude's causal effect on vote choice, as induced by campaign strategy. This is a bundled causal effect that could combine such mechanisms as changes in issue weights, changes in perceptions of candidates' platforms, or other factors.

Sides, Tesler and Vavreck (2019, 72) describe "political activation" as the process whereby voters "acquire more information about candidates" then "evaluate candidates based on

[3]The exact psychological mechanism involved in the priming process is not relevant for our purposes. It could be that political rhetoric simply makes certain attitudes more cognitively accessible. Or, voters may infer from the attention being paid to the issue that it is of great concern to general welfare. In either case, the behavioral implications in terms of the model are the same.

their long-standing political predispositions." Media coverage provides information about candidates' policy positions, their values, and their personalities. This coverage "signals to voters whether the surging candidate is 'their type,' and those whose beliefs align with the candidate's then lead the surge" (72). In the context of the 2016 election, Trump discussed issues of race, ethnicity, and religion more than was typical for candidates at the time. As a result, the authors argue, "voters' own views on these issues became more strongly related to whether they supported Trump or one of his opponents in the primary."

This account of activation involves more than simply increasing the weight that voters attach to a particular issue. Instead, it reflects that campaigns also expose voters to information about the candidates. Voters' attitudes may be "activated" not just because they infer that the issue is particularly important, but also because they learn that the candidates present distinct options on some issue.

We make this notion of activation precise through the use of two counterfactual comparisons. The first, and more straightforward, is the counterfactual choice that voters would have made had they held different opinions on the issues. In other words, it asks the question: How would the election been different, had all voters held different attitudes on a particular issue. The second is the effect that a candidate's campaign can have on the relationship between voters' attitudes and their vote choices. It asks: If a different campaign had been run, would the effect of attitudes on vote choice have been different? We define activation as this second counterfactual query, which can be expressed in the language of causal moderation effects (Bansak, 2021).

Given this formulation, we can apply standard results about identification of causal effects. In particular, the dual-counterfactual definition of activation requires strong identification assumptions: (a) that attitudes are ignorable with respect to potential outcomes; and (b) that campaigns are also ignorable with respect to attitudes and potential outcomes. The former assumption is necessary to identify the effect of attitudes on vote choice and the latter is necessary to identify the effect of campaigns on the vote choice-attitude relationship. Experimental manipulations of campaigns—e.g., vignette or laboratory experiments that randomly assign campaign messages—satisfy the second ignorability assumption. However,

the first ignorability assumption is difficult to guarantee. As a workaround, many researchers use lagged measures of attitudes to guard against reverse causality. We use our framework to study this design, and show that it entails estimation of a slightly different effect than the one that is typically targeted.

2.3 Activation as Predictive Capacity of Attitudes

A final conceptualization of activation focuses not on the causal effect of an attitude, but on its ability to predict vote choice above and beyond other variables. Activation, in this framework, refers to the possibility that campaign messaging or news media causally increased the predictive capacity of an attitude.

This prediction goal is stated in many studies of activation. For example, in a comparison of the 2012 and 2016 presidential campaigns, Hopkins (2019, 665) writes: "given President Trump's rhetoric [and] President Obama's departure . . . , it is plausible that anti-Latino prejudice could predict 2016 vote choice more strongly than 2012 vote choice." Similarly, Lajevardi, Abrajano and Diego (2019) report that because of the "campaign rhetoric toward Muslim Americans in [the 2016] presidential race, attitudes toward Muslim Americans predicted vote choice in one of the most contentious presidential elections that we have witnessed in the last several decades."[4]

This predictive goal is subtly different from the goal in our previous definition of activation, and requires weaker assumptions. Rather than the dual-ignorability assumption required when studying activation as causal moderation, assessing whether a campaign increased the *predictive* capacity of a variable requires only an ignorability assumption on the campaign. While social science theories are often stated in terms of causal effects, analyzing the capacity of an attitude to predict vote choice may be valuable in assessing campaign strategy. Evaluating an attitude's predictive power cannot isolate its effect on vote choice, but it can reveal patterns in voter behavior that underlie a campaign's success.

[4]It is possible that these researchers are using the term "prediction" more informally, in the sense of a "predictor" (covariate) in a regression model. In that case, the implied estimand may be the causal moderation effect discussed previously. Either way, it is worth drawing out the different implications of prediction per se compared to estimation of causal effects.

3 Activation In A Weighted Spatial Voting Model

Our first formulation of activation imposes a significant amount of structure on voting behavior—namely, that voters decide between candidates according to a weighted, multi-dimensional spatial voting model. This formal model makes strong assumptions on voters' decision-making process, but provides a high degree of clarity as to the meaning of activation or priming: an increase in the weight that voters attach to a given dimension. This model also encompasses the related phenomena of "learning" and "opinion change" (Lenz, 2009). The fact that the model captures these important mechanisms makes it particularly well-suited for interpreting research designs used in the literature. We relate the behavioral parameters of the model—which capture the substantive phenomena of interest—to statistical parameters estimated with data.

We particularly focus on the coefficients obtained from regressing vote choice on issue attitudes. We show that these coefficients are a product of voters' issue weights and the divergence in candidates' platforms on that issue. This result is intuitive: if voters place no weight on an issue X, then their attitude on that issue should not correlate with their vote choice. Similarly, if voters perceive there to be no difference between candidates on issue X, then their attitude on X should not correlate with their vote choice, regardless of how important that issue is. The upshot of this discussion is that cross-election changes in the relationship between vote choice and X do not yield information about the importance that voters attach to X unless the candidates' platforms are identical in both elections.

The model suggests an alternative empirical approach that focuses on voters' ratings of individual candidates—rather than comparisons between candidates—that may be more fruitful for identifying issue weights in elections. Additionally, it suggests that experimental studies of priming should seek to measure outcome variables beyond candidate ratings or vote choice, including respondents' own issue positions and their perceptions of candidates' positions.

3.1 A Spatial Model of Voting

Consider a spatial voting model with voters (indexed by i) and candidates (indexed by j), in which both have preferences defined over policies in 2-dimensional space. Voters' ideal points are denoted $\Theta_i = (\theta_i^1, \theta_i^2)$ and candidates' ideal points are denoted $\mathbf{x}_j = (x_{j,1}, x_{j,2})$, with voters choosing between candidates D and R, $j \in \{D, R\}$.

The spatial component of utility that a voter receives from candidate j is a function of the distance between her ideal point and the candidate's ideal point, where each voter has a weighting vector $\mathbf{w}_i = (w_{i,1}, w_{i,2})$ describes how much weight she attaches to each dimension. We will suppose that $w_{i,1} \in [0, 1]$ and that $w_{i,2} = 1 - w_{i,1}$. Assuming a weighted Euclidean distance metric defined by $\mathbf{w}_i$, we can write the utility voter i gets from candidate j as

$$U_i(\mathbf{x}_j; \Theta_i, \mathbf{w}_i) = -\sum_{k=1}^{2} w_{i,k}(x_{j,k} - \theta_{i,k})^2. \tag{1}$$

In deciding how to vote, voters consider the spatial component of utility described by $U_i(\cdot)$ as well as an additive, independently distributed error term for each candidate, denoted by v_{ij}. This term could correspond to valence qualities of the candidate, such as perceived competence, or other determinants of vote choice that are not related to policy positions. Voters then choose the candidate who gives them the higher utility.[5] Voters will vote for D over R if the spatial and non-spatial utility obtained from D is higher than that from R:

$$\begin{aligned} U_i(\mathbf{x}_R; \Theta_i, \mathbf{w}_i) + v_{iR} &< U_i(\mathbf{x}_D; \Theta_i, \mathbf{w}_i) + v_{iD} \\ \implies v_{iR} - v_{iD} &< U_i(\mathbf{x}_D; \Theta_i, \mathbf{w}_i) - U_i(\mathbf{x}_R; \Theta_i, \mathbf{w}_i). \end{aligned}$$

Assuming v_{ij} are each independently, normally distributed with variance $1/2$, we have a weighted form of the classic spatial voting probit model (Clinton, Jackman and Rivers, 2004):

$$\begin{aligned} P(\text{voter } i \text{ votes } D) &= \Phi\Big(U_i(\mathbf{x}_D; \Theta_i, \mathbf{w}_i) - U_i(\mathbf{x}_R; \Theta_i, \mathbf{w}_i)\Big) \\ &= \Phi\left(\sum_k w_{i,k}[(x_{R,k} - \theta_{i,k})^2 - (x_{D,k} - \theta_{i,k})^2]\right), \end{aligned} \tag{2}$$

[5] In this discussion, we assume all voters turn out. Later, we show how choices about how to analyze non-voters can bias empirical conclusions.

where $\Phi(\cdot)$ is the standard normal cdf.[6]

If we have estimates of voters' and candidates' issue-specific ideal points on the same scale then we can estimate Equation 2. Unfortunately, such data are rarely available. While a number of studies have developed measures of both politicians' and the public's ideal points on the same scale (e.g., Gerber and Lewis, 2004; Jessee, 2009; Bafumi and Herron, 2010; Bonica, 2014), these studies nearly always use one-dimensional measures of ideology. This simplification is often useful but prevents estimation of the weights attached to different issues. Perhaps not surprisingly, few studies of attitude activation attempt to explicitly adjust for the positions candidates take.

3.2 Single-Election Regressions of Vote Choice on Attitudes

In the absence of the rich data measuring voters' and candidates' policy preferences on the same scale necessary to directly estimate Equation 2, researchers often estimate simpler models that regress vote choice solely on voters' attitudes (e.g. Valentino, Hutchings and White, 2002; Reny, Collingwood and Valenzuela, 2018; Sides, Tesler and Vavreck, 2019). The spatial model outlined here provides a framework for interpreting the estimates from this reduced-form regression.

Suppose we regress vote choice on a set of issue attitudes Θ_i in a probit model:

$$P(\text{voter } i \text{ votes } D) = \Phi\left(\alpha + \beta_1\theta_{i,1} + \beta_2\theta_{i,2}\right). \tag{3}$$

The coefficients $\beta = (\beta_1, \beta_2)$ capture, roughly, the conditional correlation between vote choice and issue attitudes. The estimated regression yields predicted probabilities of voting for D among voters with different ideal points Θ_i.

To see why we cannot use the coefficients to infer the weight voters attach to dimensions, we relate the coefficients estimated in the reduced form regression in Equation 3 to the structural parameters of the voters' behavioral model given by Equation 2. To make this comparison, we make the simplifying assumption that all voters hold the same set of issue

[6]The second line follows from substitution for $U_i(\cdot)$ given in Equation 1 and some algebra. Alternative specification of the error terms yield similar expressions, such as a logit or linear probability model (McFadden, 1978; Poole and Rosenthal, 1997; Heckman and Snyder, 1997).

weights, so $\mathbf{w}_i = \mathbf{w}$ for all i.[7] Under the weighted spatial voting model with homogeneous weights, the reduced form parameters α and $\beta = (\beta_1, \beta_2)$ are functions of the structural parameters $(\Theta_i, \mathbf{w}_i, \mathbf{x}_D, \mathbf{x}_R)$:

$$\alpha = \sum_k w_k(x_{R,k}^2 - x_{D,k}^2) \qquad \text{and} \qquad \beta_k = 2w_k(x_{D,k} - x_{R,k}). \tag{4}$$

Equation 4 shows that the reduced-form probit model coefficients are functions of both the weights voters attach to the issue and the positions of the campaigns. And therefore, there are infinitely many combination of weights and candidate positions that are consistent with a particular estimated coefficient value.[8]

This result makes it difficult to interpret cross-election comparisons in terms of issue weights. For example, Sides, Tesler and Vavreck (2019) examine the relationship between racial resentment and vote choice in the 2012 and 2016 elections, finding a greater correlation in 2016 than in 2012.[9] In our notation, the empirical finding is that $\beta_{2016} > \beta_{2012}$. Given the result in Equation 4, this relationship implies that $w_{2016}(x_{D,2016} - x_{R,2016}) > w_{2012}(x_{D,2012} - x_{R,2012})$. We can see immediately that larger coefficients could be because voters attached more weight to the issue in 2016 than in 2012, or because the candidates were further apart on racial issues. In fact, it could be that voters attached *less* weight to the issue in 2016, but the candidates were so much further apart that the resulting coefficient increased from 2012 to 2016.

If we observe data on candidates' positions, then we could infer whether the weight attached to an issue has increased. Or, if we make the assumption that the difference in racial policy between Obama and Romney in 2012 was less than the difference in policy between Clinton and Trump in 2016, then we could infer that the weight voters attached to racial issues was greater in 2016 than in 2012. Alternatively, if we assume that the issue

[7] This assumption is almost certainly implausible, but relaxing it makes analysis more problematic. Rivers (1988) shows that, in general, the presence of heterogeneity in issue weights will bias estimation of average issue weights. One sufficient condition to avoid this bias is that the issue weights are uncorrelated with individuals' ideal points.

[8] In particular, with K dimensions Equation 4 has $3K - 1$ unknowns—a K-dimensional ideal point for each candidate D and R, plus a $(K-1)$-dimensional vector of issue weights—and only $K + 1$ equations.

[9] As noted above, Sides, Tesler and Vavreck's (2019) definition of activation does not appear to be solely about issue weights. We use this example here only because it provides a clear illustration of our more general point.

weights are the same in each election, then we can interpret the coefficient as the difference in policy divergence from 2012 to 2016. In any case, comparison of the reduced-form coefficients is not enough to disentangle these different potential data-generating processes. Further assumptions on candidate platforms are necessary to back out issue weights from the vote choice-attitude regression.

3.3 Recovering Issue Weights Using Measures of Approval

In much research on priming and activation, the outcome of interest is vote choice—a comparison between two candidates. However, occasionally researchers focus on approval of individual politicians. For example, Hart and Middleton (2014) consider how news media affect the correlation between overall presidential approval and respondents' views of how the president is handling select issues.

Some of the difficulties in estimating issue weights using vote choice data can be alleviated by using approval, or some other proxy for the utility a voter assigns to a candidate, as the outcome measure. Suppose we have a continuous measure of approval: for example, presidential job approval, or perhaps a feeling thermometer score for a candidate. Further, assume that this approval measure is equal to voter utility from the candidate, as given by Equation 1, plus a mean-zero error term.[10]

In this scenario, the prospects for obtaining the issue weights are more hopeful than in the vote-choice scenario. The reason is that vote choice depends on the policy positions of both candidates, while measures of approval evaluate only one candidate. In particular, note that we can rewrite utility in Equation 1 in the following terms

$$U_{i,j} = a + \sum_k \left(\delta_k \theta_{i,k} + \gamma_k \theta_{i,k}^2\right). \tag{5}$$

One can show that $\gamma_k = w_k$, $\delta_k = 2w_k x_{j,k}$, and $a = -\sum w_k x_{j,k}^2$. Thus, regressing approval on respondents' issue positions and their squares can recover issue weights, and the candidate's

[10]We consider continuous measures of utility for simplicity here. However, the basic insight would also apply to discrete measures of approval, along with a latent variable model that links (latent) continuous approval to the coarsened measured approval. Alternatively, ranking data over candidates—i.e., many binary comparisons between more than just two candidates—can also help identify issue weights. Rivers (1988) takes this latter approach, converting feeling thermometer data into ordinal rankings of candidates.

issue positions can be recovered through a simple function of the regression coefficients: $x_{j,k} = \delta_k / 2\gamma_k$.

This simple result suggests that using approval measures or candidate feeling thermometer ratings offer more promise for determining the importance that voters attach to different issues. Studies of activation or priming would be better served, therefore, by using these measures than by using vote choice data.[11]

3.4 Turnout Decisions and Selection Bias

Applied researchers studying activation often restrict their samples to those who turn out to vote—ignoring non-voters and (often) third-party voters. To study the consequences of this analysis decision, we modify the spatial voting model to include abstention. Dropping non-voters leads to post-treatment bias if voters' turnout decisions are affected by the same policy issues that influence candidate choice—making interpretation of the regression coefficients even more difficult. Under the proposed model extension, an ordered regression model recovers the correct slope coefficients.

Our simple extension to the spatial voting model is to incorporate a "calculus of voting" turnout decision (Riker and Ordeshook, 1968; Kawai, Watanabe and Toyama, 2019). We assume that voters abstain if they are close to indifferent between the candidates. In particular, suppose that voters abstain if $|(U_i(\mathbf{x}_D; \Theta_i, \mathbf{w}_i) + v_{i,D}) - (U_i(\mathbf{x}_R; \Theta_i, \mathbf{w}_i) + v_{i,R})| < c$, where c may be thought of as the cost of voting. In this case, the vote choice model follows an ordered probit (or logit) specification. Estimating such a model yields identical insights about the structural interpretation of the slope coefficients.

On the other hand, estimating a binary choice model using data only on voters who turned out can lead to badly biased estimates of the coefficients. The reason is that selection into the sample is governed by the same variables as the vote-choice decision, causing classic selection bias. To demonstrate this bias in a stylized setting, we perform a simple simulation. We suppose that voters' ideal points are over two dimensions and are drawn from a multivariate

[11] Inferences could be improved by using ratings data on multiple candidates; however, care must be taken to ensure that cross-candidate conclusions do not depend on arbitrary assumptions (such as unit-variance valence shocks) used to identify probit coefficients.

normal distribution $\Theta_i \sim$ Multivariate Normal $\left(\mathbf{0}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$ and that all voters place equal weight on each dimension, $w_1 = .5$ and $w_2 = .5$. We suppose the Democratic candidate adopts positions $\mathbf{x}_D = (0, 0)$ and the Republican adopts positions $\mathbf{x}_R = (1, 1)$. We then vary the cost from from 0, 1, and 2 and simulate turnout and voting decisions. For each cost parameter, we estimate two regressions. First, we run a probit model regressing vote choice on the two-dimensional ideal points, subsetting to simulated voters who turn out. Then, we run an ordered probit regression, coding a vote for the Democrat as -1, abstention as 0, and a vote for the Republican as 1. Because we generated the data according to the spatial voting model, the population regression coefficients are given by Equation 4. Specifically, the intercept $\alpha = 1$ and the slope coefficients $\beta_1 = -1$ and $\beta_2 = -1$.

Table 1 shows the results of this simulation, demonstrating that subsetting the analysis to just those voters who turn out biases estimation of the coefficients. To see the effect of excluding some voters, we first focus on the α, β_1, and β_2 columns to to the left of Table 1. When $c = 0$ all voters turn out and the estimated parameters are very close to the true values. But with non-zero voting costs some voters decide to not participate. And even though the other parameters of the model—particularly, voters' ideal points, issue weights, and candidate platforms—remain unchanged, we see that subsetting to only those voters who turn out creates the impression that the attitudes are activated. For example, when $c = 1$ the estimated coefficients are $\beta_1 = -1.71$ and $\beta_2 = -1.70$. And when the costs increase to 2 the coefficients are even larger in magnitude, $\hat{\beta}_1 = -2.52$ and $\hat{\beta}_2 = -2.46$.

The rightmost columns show the slope coefficients from ordered probit regressions. Here, because the statistical model takes the turnout decision into account and is correctly specified, the coefficients closely match the theoretical values across all values of the cost parameter.

This simple example shows that bias is created by subsetting to voters who turn out. This practice can create the illusion of "activation"—via an increase in the correlation between vote choice and an attitude—even when none of the underlying behavioral parameters change. In general, conditioning on turnout is a general issue of post-treatment bias (Knox,

Table 1: Bias Associated with Restricting Samples to Those Who Turned Out

	Only Voters Who Turned Out (Probit)			All Voters (Ordered Probit)	
Cost	α	β_1	β_2	β_1	β_2
0	1.00	-0.99	-0.99	-0.99	-0.99
1	1.70	-1.71	-1.70	-1.00	-1.00
2	2.53	-2.52	-2.46	-1.00	-1.01

Notes: This table presents estimated slope coefficients from a regression of vote choice on attitudes, using simulated data, across various cost parameters. The true slope parameters are $\beta_1 = \beta_2 = -1$ and the true intercept parameter is $\alpha = 1$. The left-hand panel estimates a binary probit model on those who turn out to vote, while the right-hand panel estimates an ordered probit model. The estimates of the slope parameters are severely biased when restricting the sample to those who decide to turn out.

Lowe and Mummolo, 2020; Nyhan, Skovron and Titiunik, 2017). If turnout decisions are systematically related to the attitudes that social scientists want to assess, then only examining the vote choice decisions of those who turnout to vote can create an incorrect impression of activation.

3.5 Interpreting Experimental Campaign Studies Using the Spatial Voting Model

Studies of priming often employ survey or laboratory experiments in which respondents are randomly exposed to some campaign messaging, news media, or other stimuli. To take a concrete example, Valentino, Hutchings and White (2002) show respondents campaign advertisements for the 2000 presidential election, which vary in the extent to which "racial cues [are] embedded in standard political appeals" (78). While the narration in the ads is constant across treatment conditions, the on-screen images vary. In the placebo control condition, respondents view "racially neutral visuals such as the Statue of Liberty," while in the treatment conditions, "visual racial cues are substituted for some of the neutral symbolism" (79). The authors report that racial attitudes (measured pre-treatment) are significantly more correlated with vote choice in the treatment conditions than in the control condition.

This design is useful because the random assignment of campaign messages enables un-

biased estimation of the effect of the treatment on the correlation between candidate evaluations and attitudes. However, the theoretical interpretation of the treatment effect depends on the way voters make decisions. The spatial voting framework provides one way to interpret the treatment effect: differences in the relationship between vote choice and issue attitudes across treatment conditions may reflect differences in the weights respondents attach to that issue, or they may reflect differences in perceived candidate positions. The latter portion of the treatment effect reflects the possibility that the campaign message leads respondents to update their views about the stances of the candidates—what Lenz (2009) refers to as "learning." In the Valentino, Hutchings and White (2002) example, respondents in the treatment condition might see the racialized imagery and conclude that the candidates are further apart on racial issues than they previously thought.[12]

Even under random assignment of campaign environments, standard experimental designs cannot distinguish between these two possibilities—activation and learning. The model outlined here suggests a potential improvement. Researchers interested in activation-as-issue-weights could directly measures respondents' perceptions of the candidates' platforms (ideally on the same scale as respondents' attitudes are measured). This addition would enable researchers to disentangle learning and priming effects, and directly estimate a model like the one given in Equation 2. While this approach is assumption-laden, the assumptions are transparent and the theoretical interpretation of treatment effects is clear.

Absent data on perceptions of candidates' platforms, researchers can make assumptions similar to those outlined at the bottom of Section 3.2 to interpret treatment effects as reflecting differences in issue weights. In particular, if researchers assume that the treatment does not influence evaluations of the candidates' platforms, then differences in the estimated coefficients are attributable to differences in issue weights.[13]

[12] Another concern with this particular study, and other priming studies, is that racial attitudes are measured post-treatment. If the treatment affects racial attitudes directly, then this would further complicate interpretation of the results. See Hart and Middleton (2014) for a discussion of the limitations of priming studies that measure attitudes post-treatment.

[13] Again, maintaining the assumption that the data are generated according to the weighted spatial voting model.

4 Activation as a Causal Moderation Effect

Section 3 outlines a stylized model of candidate choice, which admits a clear definition of activation as a change in the weight voters place on an issue. In that setting, it is difficult to identify the weights voters attach to issues without richer data on candidate positioning than is typically available.

The advantage of the formalization is clarity in theoretical mechanisms that could influence vote choice. The clear disadvantage is that it assumes an underlying behavioral model that might not accurately describe how voters make decisions. Indeed, key research focuses on underlying cognitive processes involved in priming, rather than a rational choice framework.[14]

In this section, we generalize the intuition from the structural approach of Section 3, relaxing the stringent assumptions about how voters make decisions. We define a reduced-form estimand for activation in terms of the causal effect of holding an attitude on vote choice and the influence of campaigns on that effect. We provide an accompanying set of assumptions that enable us to assess whether campaigns activate a particular set of attitudes.

4.1 Vote Choice as a Function of Attitudes and Campaigns

To begin, we assume that vote choices (or evaluations of candidates) Y_i are influenced both by voters' own issue attitudes, which we denote θ_i, and the campaign environment, denoted $\mathbf{x}_i$. For notational simplicity, we assume attitudes are binary, $\theta_i \in \{0, 1\}$. This allows us to define treatment effects in terms of simple contrasts.[15] We define voter i's decision—e.g., their vote choice or their evaluation of the candidates—by the potential outcomes $Y_i(\mathbf{x}_i, \theta_i)$.

The first building block of our conceptualization of activation is the *attitude average treatment effect* (AATE), which is defined as the treatment effect of holding one attitude

[14]For example, Miller and Krosnick (2000) attempt to adjudicate two explanations for priming effects seen in news media. The first, "cognitive," explanation is that seeing news media cover an issue increases the accessibility of that issue—regardless of whether, upon reasoned reflection, citizens would conclude that the issue deserves special consideration. The second, "rational," explanation is that citizens may infer that an issue is particularly important when they see news media discussing it.

[15]More generally, θ_i could be continuous, and we could define treatment effects in terms of any two values of θ_i.

relative to another, holding constant the campaign environment:

$$\text{AATE}(z_i) = E[Y_i(\mathbf{x}_i, 1) - Y_i(\mathbf{x}_i, 0) \mid \mathbf{x}_i = z_i]. \tag{6}$$

Identifying the AATE in a single election requires standard ignorability assumptions—namely, that attitudes are (conditionally) independent of the potential outcomes. Analysts often rely on an extensive set of control variables to make this assumption more plausible. However, this is a strong assumption that typically cannot be guaranteed in practice—even experimental designs may not be able to effectively manipulate attitudes. That said, for the moment we are interested in campaign effects and will assume that the AATE can be consistently estimated in any given election.

The AATE admits several theoretical interpretations, one of which is given by the spatial voting in the previous section. Under that model, the effect of manipulating attitudes depends on the weight attached to the issue (which could vary at the individual level and is not explicitly modeled here) and candidates' positions on the issue (which is implicitly included in $\mathbf{x}_i$). The fact that the AATE is a function of the campaign environment highlights that the effect of manipulating an attitude depends not only on individual factors, but also on the stances that candidates take, the state of the economy, the issues the media focuses on, and so on.

A natural question, which is prevalent in the activation literature, is how the AATE varies across elections. We represent two elections (or campaign environments) using the notation $\mathbf{x}_i'$ and $\tilde{\mathbf{x}}_i$. This estimand is the *conditional average treatment effect* (CATE), which is defined as:

$$\begin{aligned}\text{CATE}(\mathbf{x}_i', \tilde{\mathbf{x}}_i) &= \text{AATE}(\mathbf{x}_i') - \text{AATE}(\tilde{\mathbf{x}}_i) \\ &= E[Y_i(\mathbf{x}_i, 1) - Y_i(\mathbf{x}_i, 0) \mid \mathbf{x}_i = \mathbf{x}_i'] - E[Y_i(x_i, 1) - Y_i(\mathbf{x}_i, 0) \mid \mathbf{x}_i = \tilde{\mathbf{x}}_i].\end{aligned} \tag{7}$$

This estimand is non-causal: it simply asks how the treatment effect of an attitude varies across two different elections (Bansak, 2021). It is a quantity related to treatment effect heterogeneity across different observed campaign environments. The assumptions to estimate this non-causal quantity are simply that the AATE under both sets of campaign environments are identified, i.e. that we can estimate $\text{AATE}(\mathbf{x}_i')$ and $\text{AATE}(\tilde{\mathbf{x}}_i)$.

This quantity may be descriptively interesting. It is important to know if the effect of attitudes on vote choice differs across elections. However, as we stressed before, the campaign environment that $\mathbf{x}_i$ represents includes many things beyond campaign strategy per se, such as the state of the economy, the occurrence of natural disasters or pandemics, and so on. Even if the candidates in two elections were to adopt exactly the same platforms, make the same speeches, and run the same television ads, other factors in the world may cause there to be different effects of holding attitudes on vote choice.

4.2 Activation as Campaigns Changing the Effect of Attitudes on Vote Choice

While the CATE may be descriptively interesting, researchers studying activation are typically interested in the *causal* effect of campaigns. For example, Hopkins (2019, 664) writes that "political rhetoric can make people's pre-existing attitudes toward social groups more central in their support for candidates, an effect known as priming or activation." This characterization implicitly includes a counterfactual comparison: some political rhetoric can increase the effect of attitudes on vote choice, relative to other (counterfactual) rhetoric.

We formalize this notion of activation as the *average treatment moderation effect* of a particular campaign strategy $\mathbf{x}_i'$ relative to a counterfactual campaign strategy $\tilde{\mathbf{x}}_i$ (Bansak, 2021, 68). This estimand is defined as:

$$\text{ATME}(\mathbf{x}_i', \tilde{\mathbf{x}}_i) = E[Y(\mathbf{x}_i', 1) - Y(\mathbf{x}_i', 0)] - E[Y(\tilde{\mathbf{x}}_i, 1) - Y(\tilde{\mathbf{x}}_i, 0)]. \tag{8}$$

This difference between the ATME and the CATE is subtle but important. The CATE compares the effect of attitudes on vote choice across different observed campaign environments, whereas the ATME compares treatment effects of attitudes on vote choice under counterfactual campaign environments.

An example makes the distinction clear. Research has found that racial attitudes more strongly correlated with vote choice in 2016 than in prior elections (Sides, Tesler and Vavreck, 2019). Let us assume that these correlations reflect the causal effect of holding particular racial attitudes within each election. This assumption alone is sufficient to identify the

CATE—the difference in the attitude average treatment effects in the 2012 and 2016 campaigns. However, suppose we are interested in whether Trump's rhetoric was *responsible* for this difference. Here, we must imagine a counterfactual Trump campaign, in which his stances and rhetoric on racial issues were less inflammatory.[16] The simple comparison of AATEs in 2012 and 2016 is not enough to answer this question, unless we make the implausible assumption that the two campaign environments were exactly the same, but for Trump's rhetoric. Many things differed between these two elections besides Trump's rhetoric—including the candidates, the state of the economy, and the salience of different issues. Any of these factors could, in part, explain the observed CATE.

This example illustrates that attributing a change in treatment effect across elections to different campaign strategies—i.e., estimating the ATME for two campaign strategies—demands a stronger set of assumptions than is needed to identify the CATE. Bansak (2021, 70) shows that a sufficient condition is that attitudes are ignorable—as good as randomly assigned—and that campaign strategies are also ignorable.[17] Formally, these conditions are:

$$\begin{aligned} Y(\mathbf{x}_i', 1), Y(\mathbf{x}_i', 0), Y(\tilde{\mathbf{x}}_i, 1), Y(\tilde{\mathbf{x}}_i, 0), x_i \quad &\perp\!\!\!\perp \quad \theta_i \\ Y(\mathbf{x}_i', 1), Y(\mathbf{x}_i', 0), Y(\tilde{\mathbf{x}}_i, 1), Y(\tilde{\mathbf{x}}_i, 0) \quad &\perp\!\!\!\perp \quad x_i \end{aligned} \tag{9}$$

This double-ignorability assumption is strong. It requires not only that attitudes are as-if randomly assigned, but also that the campaign context is as-if randomly assigned.

Of course, experimental designs—like the Valentino, Hutchings and White (2002) study discussed above—guarantee that the campaign context is randomly assigned. Under the additional assumption that attitudes are also randomly assigned, the ATME is identified in this sort of design. In observational settings, this ignorability assumption is more difficult to maintain. Candidates may select campaign rhetoric based on their impression of voters' attitudes or based on how voters will respond to different counterfactual campaign strategies. Both are violations of the ignorability assumptions.

[16] The fact that we must measure the effect of Trump's rhetoric relative to some specific counterfactual is implicit in the definition of the ATME: Equation 8 involves a comparison between two particular campaign environments, $\mathbf{x}_i'$ and $\tilde{\mathbf{x}}_i$.

[17] These ignorability assumptions could also be made conditional on covariates, in which case the formal notation would have to be modified to state conditional independence.

4.3 Panel Design

In the prior sections, we have assumed that individuals' attitudes are randomly assigned. This simplifying assumption allowed us to focus on campaign-level effects, rather than the thorny issue of estimating the causal effect of attitudes on candidate evaluations. In real applications, however, this assumption is crucial—and often implausible. Individuals often adjust their attitudes to reflect the policy positions of their preferred candidates (Barber and Pope, 2019; Lenz, 2013). This is worrisome, because it reverses the causal ordering that we had assumed in the prior section: that attitudes affect vote choice. If, instead, voters "follow the leader" then their vote choice affects their attitudes. In a cross-sectional regression, it is impossible to distinguish between these two possibilities. And so a simple regression may indicate that an attitude is strongly related to an outcome, even though that attitude exerts no *causal* effect on vote choice.

A common strategy to ensure respondents don't change their attitudes to align with their candidate's position is to use panel data (Hopkins, 2019; Sides, Tesler and Vavreck, 2019; Mason, Wronski and Kane, 2021). The intuition is that it is impossible for a candidate's position in the future to affect the voter's position in the past. So, the bias from "following the leader" is removed. Using this lagged attitude measure, researchers then estimate the correlation between vote choice and the attitude—i.e., estimate the attitude average treatment effect (AATE) defined above.

In this section, we analyze the assumptions necessary for this research strategy to identify the AATE. We find that while this strategy avoids the threat of "follow-the-leader" bias, additional assumptions are required to interpret the AATE using lagged attitudes as identical to the AATE using contemporaneous attitudes.

We slightly change the notation introduced above, adding time subscripts. We denote individual i's attitude in year with θ_{it}. Similarly, we denote the campaign environment that i is exposed to at time t with $\mathbf{x}_{it}$. Finally, the potential outcomes—reflecting vote choice or candidate evaluations—are also indexed by time.

The typical regression with lagged attitudes attempts to estimate the effect of holding

an attitude in time $t-1$ on vote choice in time t. We call this estimand the *lagged attitude average treatment effect* (LAATE), and it is directly analogous to the AATE in Equation 6:

$$\text{LAATE}_t(\mathbf{x}_{it}) = E[Y_{it}(\mathbf{x}_{it}, \theta_{it-1} = 1) - Y_{it}(\mathbf{x}_{it}, \theta_{it-1} = 0)]. \tag{10}$$

This quantity is identified under a standard ignorability assumption on lagged attitudes θ_{it-1}, which may be more plausible than an ignorability assumption on contemporaneous attitudes θ_{it}.

The LAATE may be an important quantity to estimate if one is interested in the long-run development of political behavior. But it differs in important ways from the AATE, i.e. the effect of contemporaneous attitudes. In particular, the LAATE picks up on all the causal consequences of holding a particular attitude in the prior election, but before the current election. If the main object of inquiry is the effect of current attitudes on vote choice in the current election, then the LAATE may be a poor stand-in for the AATE.

An example makes these concerns concrete. In the example, we suppose that voting for a party in one election exerts a causal effect on voting for the same party in the subsequent election.[18] Suppose that we randomly assign individuals to have either high ($\theta_{it-1} = 1$) or low ($\theta_{it} = 0$) racial resentment in time $t-1$. Further, suppose that 60% of high-resentment voters choose the Republican ($E[Y_{it-1}(\mathbf{x}_{it-1}, 1) = 0.6$), while 35% of the low-resentment voters choose the Republican ($E[Y_{it-1}(\mathbf{x}_{it-1}, 0) = 0.35$). This implies an attitude average treatment effect in election $t-1$ of 0.25.

But now, suppose that each voter's choice in the next election, in time t, also depends on their previous vote choice. Specifically, we assume that the choice voters made in the prior election makes them more likely to support the same party's candidate in the next election. We can capture this mechanism with a new potential outcome for time t, which includes the campaign environment, contemporaneous attitudes (which we assume were randomly assigned in the previous election), and past vote choice: $Y_{it}(\mathbf{x}_{it}, \theta_{it}, y_{it-1})$. Suppose that we have the following averages in the population:

[18] A simple partisan identity mechanism could be at play here, in which voting for a party strengthens one's attachment to the party. Another mechanism could be "follow-the-leader"—after voting for a candidate, voters may update their views on a range of policy opinions to align with that candidate.

- 90% of high-resentment individuals who voted for the Republican in $t-1$ vote for the Republican in t: $E[Y_{it}(\mathbf{x}_{it}, \theta_{it} = 1, y_{it-1} = 1) = 0.9$
- 65% of low-resentment individuals who voted for the Republican in $t-1$ vote for the Republican in t: $E[Y_{it}(\mathbf{x}_{it}, \theta_{it} = 0, y_{it-1} = 1) = 0.65$
- 35% of high-resentment individuals who voted for the Democrat in $t-1$ vote for the Republican in t: $E[Y_{it}(\mathbf{x}_{it}, \theta_{it} = 1, y_{it-1} = 0) = 0.35$
- 10% of low-resentment individuals who voted for the Democrat in $t-1$ vote for the Republican in t: $E[Y_{it}(\mathbf{x}_{it}, \theta_{it} = 0, y_{it-1} = 0) = 0.1$

These sets of potential outcomes imply that the treatment effect of contemporaneous resentment is 0.25: within each stratum defined by previous vote choice, high-resentment individuals are 25 percentage points more likely to vote for the Republican candidate. If we re-randomized attitudes at time t, we would obtain this estimate of the attitude average treatment effect. However, if we calculate the lagged attitude average treatment effect, using the outcome at time t and assuming attitudes are randomly assigned at time $t-1$, we obtain a much larger estimate:

$$\begin{aligned}
\text{LAATE}_t(\mathbf{x}_{it}) &= E[Y_{it}(\mathbf{x}_{it}, \theta_{it-1} = 1) - Y_{it}(\mathbf{x}_{it}, \theta_{it-1} = 0)] \\
&= (0.9 \times 0.6 + 0.35 \times 0.4) - (0.65 \times 0.35 + 0.1 \times 0.65) \\
&= 0.3875.
\end{aligned}$$

The lagged attitude average treatment effect gives the impression that attitudes were particularly activated in election t. However, by construction, the attitudes at time t and the attitudes at time $t-1$ would have produced the same treatment effect on outcomes measured at time of the attitudes' random assignment. In this example, we could potentially estimate a direct effect of θ_{it-1} on vote choice in time t by conditioning on prior vote choice. But such a strategy relies on accounting for all possible mediators from prior attitudes to current vote choice, aside from current attitudes.

Additionally, the challenges that come along with attributing differences in AATEs to the campaigns also carry over when using lagged attitudes. In order to attributed differences in lagged attitude average treatment effects across elections to the campaign environment, we must make an ignorability assumption about the campaign. Thus, using lagged attitudes may help with the assumption that attitudes are ignorable, but it does nothing to strengthen the assumption that campaign strategy is ignorable.

Working with lagged attitudes, then, represents a trade off. On the one hand, working with attitudes from prior elections bolsters the selection on observables assumption. On the other hand, it estimates a different causal effect that may not correspond to the effect of interest. Regardless of the time period that attitudes are measured, we need stronger assumptions than random assignment of attitudes in order to estimate activation effects of campaigns.

4.4 Turnout in the Causal Moderation Framework

In this section, we have ignored turnout decisions for expository purposes. But just as before, subsetting only to voters who have turned out to vote creates a problem of post-treatment bias. We therefore recommend explicitly including turnout decisions in activation analyses.

A simple way to include turnout is to redefine the dependent variable in two-party elections is as follows:

$$Y_{it} = \begin{cases} -1 & \text{if } i \text{ voted for the Democrat in election } t \\ 0 & \text{if abstained or vote for 3rd party} \\ 1 & \text{if voted for the Republican} \end{cases}$$

This redefinition has a simple interpretation in terms of the net votes cast for the Republican (Grimmer and Marble, 2019): the average of this variable is simply the proportion of the population who voted for the Republican minus the proportion who voted for the Democrat.

An alternative is to explicitly study support for a candidate, coding the variable as 1 if a respondent supports the candidate and 0 otherwise (including if they do not turn out to vote). The average of this variable is the share of the population who supports the candidate. Either of these dependent variables avoids the post-treatment bias induced by conditioning on turnout.

5 Activation as Prediction

Rather than focusing on features related to the causal effect of attitudes on vote choice—as do both activation-as-issue-weights and activation-as-causal-moderation—we might instead ask how well an attitude can *predict* vote choice, and how that relationship varies across elections. This formulation is more modest than the first two, in that prediction does not require the stringent ignorability assumption on attitudes that the first two formulations require. However, in order to make statements about the causal effect of campaigns, we still need an ignorability assumption on the campaign environment.

Before proceeding, it is worth noting that the distinction between prediction and causation is important. In the prediction version of activation, we are not asking whether holding a particular attitude causes one to vote in a particular way. Instead, we are merely asking whether knowing someone's attitude gives us information about how they voted. A highly predictive variable does not imply that manipulating that variable would change one's vote choice. While prediction may be important for practitioners seeking to identify potential supporters and describe patterns in voter behavior, it cannot speak to underlying *causes* of vote choice.

5.1 Measuring Predictive Performance

To begin, we lay out a general formulation for studying prediction that is based on variable importance measures (VIMs) (Wei, Lu and Song, 2015). VIMs are a set of tools used in machine learning to quantify how much better a prediction a model can make if it has knowledge of a certain variable, relative to the prediction it makes without knowing that variable. Formally, this is stated in terms of a loss function $L(y, \hat{y})$, which gives higher values when the prediction $\hat{y}$ is farther from the true value y. An example of a commonly used loss function is mean squared error, defined as: $MSE = \frac{1}{n}\sum(\hat{y}_i - y)^2$.

Define our prediction using knowledge of variables Z_i and X_i (which may be a vector) as $\hat{E}[y \mid Z_i, X_i]$. In general, variable importance for some variable Z_i is defined as the expected increase in the loss function when the predictions are made based on Z_i (in addition to other

covariates X_i), relative to when they are made without knowledge of Z_i:

$$\text{VIM}(Z_i) = E[L(y_i, \hat{E}[y_i \mid X_i])] - E[L(y_i, \hat{E}[y_i \mid X_i, Z_i])]. \tag{11}$$

This formulation of variable importance stands in contrast to the comparison of linear regression coefficients. In a linear model, the magnitude of standardized regression coefficients correspond to variable importance only under restrictive conditions.[19] As a result, it is typically not possible to gauge the predictive importance of a variable in a linear regression model solely from the regression coefficient.

This discussion also suggests that linear regression models may not be the best tools for measuring the predictive importance of a variable. Linear models impose strict functional form assumptions. More flexible methods—such as random forests, Bayesian additive regression trees, and so on—that can automatically detect interactions and nonlinearities may be better suited to judging the importance of a variable for prediction.

5.2 Campaigns' Effect on Predictive Performance

The goal of activation analysis is typically to study the effect of campaign and media. In the predictive framework, we can formalize this by introducing potential outcomes for prediction. Let us define the potential outcomes $Y_t(\mathbf{x})$ as the predictive performance of a variable of interest (say, racial resentment, attitudes toward immigration, and so on) under the campaign environment defined by $\mathbf{x}$. That is, the outcomes Y_t is a variable importance measure that assesses how important a variable is for predicting voting behavior.

We are then interested in the causal effect of campaigns on the predictive performance. We call our activation estimand the campaign treatment effect on prediction (CTEP):

$$\text{CTEP}_t(\mathbf{x}', \tilde{\mathbf{x}}) = E[Y_t(\mathbf{x}') - Y_t(\tilde{\mathbf{x}})]. \tag{12}$$

As in the previous section, treatment effects here are defined in terms of contrasts between two potential campaign environments, $\mathbf{x}'$ and $\tilde{\mathbf{x}}$. In order to identify this effect, we need a

[19] These conditions are that the linear model is the correctly specified and that the included variables are uncorrelated with each other (Wei, Lu and Song, 2015). This latter condition, especially, is violated in virtual every social science study.

standard ignorability assumption on the campaign environment; formally, $(Y_t(\mathbf{x}'), Y_t(\tilde{\mathbf{x}})) \perp\!\!\!\perp \mathbf{x}$. Substantively, this assumption means that candidates and news media do not strategically alter their behavior based on the potential predictive capacity of different variables. This is a difficult assumption to make, as campaigns actively attempt to identify potential voters using observable data.

There may be a non-zero CTEP regardless of whether the variable of interest has a causal effect on voting. Moreover, measures of variable importance also depend on the other variables considered in the model. To illustrate, suppose a mayoral campaign in a highly segregated city effectively targets Black voters for mobilization and persuasion. Due to the correlation between residential location and race, analysts who do not consider race directly in their prediction model may conclude that zip code is a highly important variable for predicting vote choice—even though it does not exert a causal effect on voting. However, if the analyst expands their predictive model to include race directly, they would come to a different conclusion: zip code is not as important as previously thought, since it largely proxied for race. Once race is included, the predictive importance of residential location is diminished.

Finally, it is worth noting that we have not added a subscript i to the potential outcomes as we had before. The reason is that prediction is a population-level task. In our case, the goal is to assess how well, in the population of eligible voters, knowledge of Z_i allows us to predict vote choice. The potential outcomes refer to population-level measures of variable importance under different counterfactual campaigns. In contrast, in the previous section, we were interested in individual-level vote choice and vote choice under counterfactual campaign environments. This allowed us to express individual-level potential outcomes.

In sum, activation can be well-defined even in a predictive framework. Scholars interested in how predictive of vote choice a variable is should employ variable importance measures, which are specifically tailored to the task. In this framework, we can define activation as the change in predictive performance induced by a campaign environment, relative to some counterfactual campaign environment. Inferences about activation thus require an ignorability assumption on the campaign strategy, but do not require the strict individual-

level ignorability assumption necessary to study activation as a causal moderation effect.

6 Conclusion

Researchers interested in elections and political behavior more broadly often turn to "priming" or "activation" as a framework for understanding the effects of political communication. According to the activation hypothesis, when campaigns and news media focus on some particular issues, it causes citizens to bring their evaluation of candidates in line with their pre-existing attitudes on those issues. Despite a plethora of studies employing this framework, the literature is often unclear about what exactly the priming or activation estimands are.

In this paper, we propose three formalizations of activation based on different notions in the literature. We then use the formalizations to study the assumptions necessary to estimate activation effects. In the first formulation, we embed activation in a multidimensional spatial model of voting. In the model, activation is defined as an increase in the weight voters attach to some issue in making their voting decision. Using this model, we show the difficulties with identifying issue weights in a standard regression framework. Comparisons of regression coefficients across elections do not correspond to comparisons of issue weights except under strong assumptions. We propose an alternative dependent variable that allows for identification of issue weights under the assumptions of the model.

The second formalization is a reduced-form generalization of the first. Instead of focusing on issue weights in a formal model, we instead focus on the causal effect of attitudes on vote choice. We then define activation as a causal moderation effect: a campaign activates an attitude if it causes an increase in the causal effect of that attitude on vote choice, relative to some counterfactual campaign. We show that in order to identify this estimand, we must assume that both attitudes and campaign strategies are as-if randomly assigned.

Finally, we study activation in a predictive framework. Activation here is defined as a campaign causing an increase in the importance of a variable for predicting vote choice. We recommend using measures of variable importance to study predictive performance. In

this version of activation, only an assumption about as-if random assignment of campaigns is necessary—loosening the assumption about random assignment of attitudes necessary to identify causal moderation effects. While the assumptions to identify this estimand are weaker, changes in predictive capacity of a variable are less theoretically interpretable than changes in the causal effect of a variable.

An overarching takeaway is that campaign effects are very difficult to study using observational data. Comparisons across elections are typically insufficient to identify activation effects, as the campaign strategy ignorability assumption is nearly always violated. Lab and survey experiments guarantee the assumption of campaign strategy ignorability, solving some of the problems of observational studies. Studies that exploit random variation in campaign intensity—for example, due to the geography of media markets or field experiments (Gerber et al., 2011; Krasno and Green, 2008; Martin, 2020)—provide another potential mechanism to identify campaign effects. However, if we conceptualize activation as related to the causal effect of attitudes, an additional assumption about the ignorability of attitudes is required.

By providing several rigorous definitions of priming and activation, we hope to provide a framework for scholars studying these important issues. Our definitions should help scholars articulate exactly what they are studying while also spurring innovations in research design that enable more credible estimates of activation effects.

References

Bafumi, Joseph and Michael C. Herron. 2010. "Leapfrog Representation and Extremism: A Study of American Voters and Their Members in Congress." *American Political Science Review* 104(03):519–542.
URL: *http://www.journals.cambridge.org/abstract_S0003055410000316*

Bansak, Kirk. 2021. "Estimating Causal Moderation Effects with Randomized Treatments and Non-Randomized Moderators." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 184:65–86.
URL: *https://arxiv.org/pdf/1710.02954.pdf*

Barber, Michael and Jeremy C Pope. 2019. "Does Party Trump Ideology? Disentangling Party and Ideology in America." *American Political Science Review* 113(1):38–54.

Berger, Jonah, Marc Meredith and S. Christian Wheeler. 2008. "Contextual Priming: Where People Vote Affects How They Vote." *Proceedings of the National Academy of Sciences* 105(26):8846–8849.

Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science* 58(2):367–386.

Campbell, Angus, Phillip Converse, Warren Miller and Donald Stokes. 1960. *The American Voter*. Chicago: Chicago University Press.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2):355–370.

Gerber, Alan S., James G. Gimpel, Donald P. Green and Daron R. Shaw. 2011. "How Large and Long-Lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment." *American Political Science Review* 105(1):135–150.

Gerber, Elisabeth R. and Jeffrey B. Lewis. 2004. "Beyond the Median: Voter Preferences, District Heterogeneity, and Political Representation." *Journal of Political Economy* 112(6):1364–1383.

Grimmer, Justin and William Marble. 2019. "Who Put Trump in the White House? Explaining the Contribution of Voting Blocs to Trump's Victory.".
URL: *https://williammarble.co/docs/vb.pdf*

Hart, Austin and Joel A. Middleton. 2014. "Priming under fire: Reverse causality and the classic media priming hypothesis." *Journal of Politics* 76(2):581–592.

Heckman, James J. and James M. Snyder. 1997. "Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators." *The RAND Journal of Economics* 28(0):S142–S189.

Hopkins, Daniel J. 2019. "The Activation of Prejudice and Presidential Voting: Panel Evidence from the 2016 U.S. Election." *Political Behavior* .
URL: *https://doi.org/10.1007/s11109-019-09567-4*

Hutchings, Vincent L. and Ashley E. Jardina. 2009. "Experiments on Racial Priming in Political Campaigns." *Annual Review of Political Science* 12(1):397–402.
URL: *https://doi.org/10.1146/annurev.polisci.12.060107.154208*

Iyengar, Shanto and Donald Kinder. 1987. *News That Matters*. Chicago: University of Chicago Press.

Jessee, Stephen A. 2009. "Spatial Voting in the 2004 Presidential Election." *American Political Science Review* 103(01):59.

Kawai, Kei, Yasutora Watanabe and Yuta Toyama. 2019. "Voter Turnout and Preferences Aggregation." *American Economic Journal: Microeconomics* (Forthcoming).

Knox, Dean, Will Lowe and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *American Political Science Review* pp. 1–19.

Krasno, Jonathan S. and Donald P. Green. 2008. "Do Televised Presidential Ads Increase Voter Turnout? Evidence from a Natural Experiment." *Journal of Politics* 70(1):245–261.

Lajevardi, Nazita, Marisa Abrajano and San Diego. 2019. "How Negative Sentiment toward Muslim Americans." *Journal of Politics* 81(1):296–302.

Lazarsfeld, Paul F, Bernard Berelson and Hazel Gaudet. 1948. *The People's Choice*. Columbia University Press.

Lenz, Gabriel S. 2009. "Learning and Opinion Change, Not Priming: Reconsidering the Priming Hypothesis." *American Journal of Political Science* 53(4):821–837.

Lenz, Gabriel S. 2013. *Follow the Leader? How Voters Respond to Politicians' Policies and Performance*. Chicago: University of Chicago Press.

Martin, Gregory J. 2020. "The Informational Content of Campaign Advertising.".

Mason, Lilliana, Julie Wronski and John V. Kane. 2021. "Activating Animus: The Uniquely Social Roots of Trump Support." *American Political Science Review* pp. 1–9.

McFadden, Daniel. 1978. "Modeling the Choice of Residential Location." *Transportation Research Record* (673):72–77.
URL: *http://onlinepubs.trb.org/Onlinepubs/trr/1978/673/673-012.pdf*

Mendelberg, Tali. 2001. *The Race Card: Campaign Strategy, Implicity Messages, and the Norm of Equality*. Princeton, N.J.: Princeton University Press.

Miller, Joanne M. and Jon A. Krosnick. 2000. "News Media Impact on the Ingredients of Presidential Evaluations: Politically Knowledgeable Citizens Are Guided by a Trusted Source." *American Journal of Political Science* 44(2):301.

Molden, Daniel C. 2014. "Understanding Priming Effects in Social Psychology: What is "Social Priming" and How Does It Occur?" *Social cognition* 32(Supplement):1–11.

Nyhan, Brendan, Christopher Skovron and Rocío Titiunik. 2017. "Differential Registration Bias in Voter File Data: A Sensitivity Analysis Approach." *American Journal of Political Science* 61(3):744–760.

Poole, Keith T. and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. Oxford University Press.

Reny, Tyler, Loren Collingwood and Ali Valenzuela. 2018. "Vote Switching in the 2016 Election: Racial and Immigration Attitudes, Not Economics, Explains Shifts in White Voting." *Public Opinion Quarterly* (Forthcoming).

Riker, William H. and Peter C. Ordeshook. 1968. "A Theory of the Calculus of Voting." *American Political Science Review* 62(1):25–42.

Rivers, Douglas. 1988. "Heterogeneity in Models of Electoral Choice." *American Journal of Political Science* 32(3):737–757.

Sides, John, Michael Tesler and Lynn Vavreck. 2019. *Identity Crisis: The 2016 Presidential Campaign and the Battle for the Meaning of America*. Princeton, N.J.: Princeton University Press.

Tesler, Michael. 2012. "The Spillover of Racialization into Health Care: How President Obama Polarized Public Opinion by Racial Attitudes and Race." *American Journal of Political Science* 56(3):690–704.

Tesler, Michael. 2015. "Priming Predispositions and Changing Policy Positions: An Account of When Mass Opinion Is Primed or Changed." *American Journal of Political Science* 59(4):806–824.

Valentino, Nicholas A., Vincent L. Hutchings and Ismail K. White. 2002. "Cues That Matter: How Political Ads Prime Racial Attitudes During Campaigns." *American Political Science Review* 96(1):75–90.

Wei, Pengfei, Zhenzhou Lu and Jingwen Song. 2015. "Variable importance analysis: a comprehensive review." *Reliability Engineering & System Safety* 142:399–432.