

# Analyzing Candidates' Ideological Messaging Throughout the Electoral Cycle

BENJAMIN WITTENBRINK\*

DEPARTMENT OF ECONOMICS, STANFORD UNIVERSITY

ADVISED BY MATTHEW GENTZKOW

MAY 4, 2022

## Abstract:

Conventional wisdom suggests that primary elections attract a loyal base of partisans, and consequently, political candidates must take more extreme positions to secure the nomination before moderating for the general electorate. Yet, the academic literature contains little empirical evidence on candidate positioning over the electoral process. I address this gap by using congressional candidates' tweets to quantify ideological rhetoric during the 2020 election using three different approaches. First, I adopt a data-driven approach to select the most partisan bigrams and specify a multinomial model of speech; second, I use a theoretically-derived dictionary to measure the frequency of moral values associated with political convictions; finally, I specify a natural language model using a deep learning architecture. I provide one of the first empirical analyses on the evolution of candidates' ideological messaging over the entire election cycle. I find statistically significant evidence of moderation among Republican but not Democratic candidates, with mixed support for greater movement in competitive general elections. I conclude that Republicans likely face stronger incentives than Democrats to employ extreme rhetoric in primaries and thus to moderate in general elections.

**Keywords:** political communication, ideology, elections, text analysis, machine learning, social media

---

\* Wittenbrink: witten@stanford.edu. A special thank you to Prof. Matthew Gentzkow for being an incredibly supportive and generous mentor, without whom this project would not have been possible. I am very grateful to Prof. Marcelo Clerici-Arias for organizing the honors thesis program. I would also like to thank Prof. Andrew Hall, Prof. Justin Grimmer, Levi Boxell, Cody Cook, Nina Buchmann, and my peers in the economics thesis cohort for their valuable help and feedback. Finally, I thank my parents for their unwavering support. All mistakes in this thesis are my own.

## 1. Introduction

Primary elections in the United States are believed to have a polarizing effect on elections (Burden, 2001, 2004; Fiorina and Levendusky, 2006; Brady, Han and Pope, 2007; Fiorina, Abrams and Pope, 2010; Hall and Snyder, 2015; Jacobson and Carson, 2019). In particular, primary elections are thought to reward more extreme policy positions and rhetoric, as they tend to attract a loyal base of highly-engaged partisan voters and party-insiders. In recent years, the ideological orientation of these groups is thought to have further diverged (Heaney et al., 2012; Pew, 2014; Gentzkow, 2016), spurred on by the increasing polarization of the American media ecosystem (Martin and Yurukoglu, 2017; Tucker et al., 2018) as well as the elimination of institutional safeguards.<sup>1</sup> The success of Donald Trump, perceived by many to be the most extreme candidate in the 2016 Republican presidential primaries over many establishment favorites, epitomizes this view of primary elections, though similar anecdotes also exist in the arena of congressional campaigns – for example, all but one of the Senators voting against the 2020 presidential electoral certification took office in the last five years.<sup>2</sup>

While the increasing polarization of American partisans is well-documented, the political science literature contains limited empirical information on the evolution of candidate positions throughout the election cycle. For instance, Hall and Snyder (2015) note that “political science research often has surprisingly little to say about the *overall* electoral process,” instead tending to focus on the general or primary elections separately or on easier populations to measure, such as incumbents, for whom congressional roll-call behavior is available. The rare studies that do analyze ideology throughout the entire electoral process and among non-incumbents suffer from sample sizes that are “too small and/or unrepresentative to be useful” (Hall and Snyder, 2015). For example, in a study of more than 2,000 candidates for the House using responses for Project Votesmart’s National Political Awareness Test, Rogowski (2013) reports that it was only possible to calculate the ideological positions for 190 candidates.

The primary reason for this void in the literature is that data on congressional candidates,

---

<sup>1</sup>As an example, both parties have reduced the influence of superdelegates. These delegates represent party insiders and are generally perceived to have a moderating presence on the selection process. The negligible role of superdelegates in the Republican primary system is often credited as aiding Trump’s rise from party outsider to nominee.

<sup>2</sup>These are Senators Ted Cruz (R-TX, first taking office in 2013), Josh Hawley (R-MO, 2019), Cindy Hyde-Smith (R-MS, 2018), John Kennedy (R-LA, 2017), Cynthia Lummis (R-WY, 2021), Roger Marshall (R-KS, 2021), Rick Scott (R-FL, 2019), and Tommy Tuberville (R-AL, 2021).

especially those that lose and thus never take office, is sparse. One of the main obstacles is the lack of any comprehensive, quantitative measure of partisanship. The literature has largely relied on roll-call voting scores (DW-NOMINATE) for congressional candidates, which presents significant complications when extending the analysis to primary candidates – the vast majority of whom will never serve in Congress. As an alternative, studies have used donation data from the Federal Election Commission (FEC), estimating the political leaning of donors and then using the weighted average of donors to a candidate to calculate their ideology (Bonica, 2013, 2014; McCarty, Poole and Rosenthal, 2016). While these donation-based estimates largely coincide with DW-Nominate ratings for Congress members, they are not well suited for comparison between the primary and the general, as the composition of donors is intrinsically different in the two periods. For example, partisans will likely coalesce around their party nominee in the general regardless of ideology while independents will chose a side. As a result, this measure will inherently indicate movement from the primary to the general, absent any change in actual ideological rhetoric or positioning.

Yet, there is substantial theoretical work in the literature arguing that candidates pander to their base in the primary in order to secure the nomination and then proceed to moderate for the general. (Cox, 1990; Burden, 2001; Hummel, 2010; Agranov, 2016). This dynamic is often termed the post-primary moderation effect. For instance, a simple two-period extension of the Downsian model with myopic voters, which I detail in the next section, would suggest that candidates will strategically manipulate their ideological messaging to capture the median voter in each period, moderating their positions from the primary to the general. However, such a model presupposes that the primary electorate is naive, unaware that the candidate will change their position in the next period. In addition, there are clear challenges to such movement: candidates may not be able to shift the voters' perception of them and voters may punish candidates for "flip-flopping," particularly if these shifts are high-profile. More nuanced models considering these flip-flopping and visibility costs and incorporating signalling as a mechanism for candidates to reveal their true type still provide support for the post-primary moderation effect, though the magnitude of this shift is governed by these costs. Thus, candidates may choose to alter their rhetoric only when the benefits exceed the consequences – for example, in especially competitive elections, where the marginal voter is near the center.

In this paper, I address this gap in the empirical literature by using congressional candidates'

tweets during the 2020 election to quantify the evolution of ideological rhetoric throughout the election cycle in order to present quantitative evidence on the post-primary moderation effect. This focus on actual candidate speech has many important benefits. First, the emphasis on text data enables the study of all candidates without a voting record, even those that lose in the primary or general. Moreover, social media – and especially Twitter – is an essential tool for direct political communication with voters. Consequently, this focus provides an opportunity to analyze ideological rhetoric in order to elucidate how candidates are attempting to characterize their political position to voters. Finally, the high-frequency nature of tweeting behavior provides a continuous snapshot of a candidate's rhetoric, which in turn enables a fine-grained temporal analysis.

In order to translate the raw Twitter data obtained for the congressional candidates into estimates of ideological extremity, I employ three different approaches. First, I take a data-driven approach to identify the most partisan ideological rhetoric using the frequencies with which candidates invoke phrases. Based on these counts, I then specify a Multinomial Inverse Regression (MNIR) to build a model of speech that predicts partisanship (Taddy, 2013). Second, I take a theoretically-derived approach, using an externally validated and verified set of keywords to obtain the relative frequencies with which candidates invoke universalist and communitarian values, a measure that has been shown to be highly correlated to ideology (Haidt and Joseph, 2004; Enke, 2020). Finally, I specify a natural language model using a deep learning approach based on a pre-trained neural network with a transformer architecture (RoBERTa), which I fine-tune on the task of ideological prediction (Liu et al., 2019). This method allows me to capture ideology at a considerably finer level of detail than the previous methods. All of these approaches create a link between speech and quantitative measures of political ideology and yield a single continuous outcome variable, representing the ideological positioning of the candidate. With these predictions, I am able to construct a novel dataset consisting of almost every congressional candidate in the 2020 election cycle with monthly estimates of the extremity of their ideological rhetoric.

To validate the obtained measures, I assess how the MNIR and RoBERTa models perform in-sample on members of the House as well as out-of-sample on the set of Senators from the 116th Congress. These samples are quite similar, and consequently, both models perform quite well. The correlation coefficients between the MNIR and RoBERTa predictions and the true DW-Nominate 1 scores are 0.99 and 0.97 in-sample, and 0.96 and 0.93 out-of-sample. Although there do not exist

analogous gold-standard labels for the candidate sample, I compare the average of my scores for candidates that won their election and assumed office with their subsequent voting behavior in Congress. In particular, I regress the future DW-Nominate scores against my predicted scores and obtain a correlation coefficient using the fitted estimates of 0.94 for MNIR and 0.92 for RoBERTa. Thus, the obtained ideology measure appears to be well-founded and to generalize to the candidate sample. In addition, qualitatively, the bigrams with the greatest ideological lift according to the MNIR model appear to match the traditional priorities of the two parties – such as “protect\_unborn” and “less\_government” among Republicans and “free\_school” and “climate\_science” for Democrats.

With this dataset, I am able to provide one of the first empirical analyses on the evolution of congressional candidates’ ideological positioning over the course of the electoral process. My interest is in both documenting whether a shift in messaging occurs and exploring the nature and possible reasons for that shift. To this end, I first specify an event study to visualize the movement of my ideological predictions throughout the electoral cycle. Because of the asynchrony in the timing of state-level primary elections, I must estimate these specifications on subsamples of my data. Consequently, I also estimate a difference-in-difference model of moderation between the primary and general elections using my full sample. Finally, I consider the possibility of differential effects according to candidate- and race-specific attributes. In particular, I introduce interaction terms with indicators for incumbency status as well as for primary and general election competitiveness to form triple difference-in-difference models.

With the event study design, I provide visual evidence of a gradual moderating ideological shift in rhetoric among Republican candidates in early spring of 2020, just a couple months before the average primary election. This observed moderation is then sustained for the entirety of the remaining election cycle. I do not observe any systematic effect among Democratic candidates. The results of the difference-in-difference model on the full sample confirm statistically significant ideological moderation over the course of the campaign among Republican but not Democratic candidates. This finding is robust to the inclusion of candidate- and race-specific characteristics as well as candidate fixed effects. On average, I estimate Republican candidates to moderate by 0.058 (MNIR) and 0.068 (RoBERTa) DW-Nominate 1 points between the primary and the general, roughly equivalent to half of a standard deviation in the DW-Nominate distribution of congressional Republicans. This is akin to moving from the median Republican Senator Mike Crapo

from Idaho, estimated at 0.51, to a slightly more moderate Senator like Cory Gardner of Colorado (0.45) or Thom Tillis of North Carolina (0.43). Additionally, MFD estimates that Republican candidates moderate their relative usage of communalist rhetoric by 0.097 points. In contrast, no significant effects are found for Democratic candidates across all methodologies. However, I do estimate Democratic candidates to be substantially more moderate throughout the election than Republicans. For example, the MNIR model estimates the average Democrat and Republican in a 50% Trump district to lie at -0.28 and 0.47 during the primary, respectively, and then at -0.28 and 0.41 in the general.

Furthermore, I provide mixed evidence for heterogeneity in the evolution of ideological rhetoric based on general election competitiveness. According to the MNIR model, Republican candidates in competitive generals moderate substantially more than those running in uncompetitive races. In particular, using an ex-ante and an ex-post indicator for competitiveness, I find the difference to be 0.109 and 0.138 points, significant at the 5% and 1% level, respectively. However, RoBERTa and MFD do not show any significant difference between the two samples. In addition, all three methodologies do not find any significant differences among Democrats. Finally, I do not find significant differences among Republican or Democratic candidates by incumbency status or primary election competitiveness. One notable limitation of this analysis is that the sample sizes for competitive elections are extremely small, with 136 and 76 candidates for my two general election indicators and only 46 for my primary indicator.

Generalizing across three conceptually and methodologically distinct models, my results show a systematic moderating trend among Republican candidates across all congressional races consistent with the post-primary moderation hypothesis. The asymmetric moderation observed among Republicans in conjunction with recent literature documenting that Republican partisans have become more ideologically extreme than Democrats ([McCarty, Poole and Rosenthal, 2016](#); [Grossmann and Hopkins, 2015](#); [Lewis et al., 2022](#)) suggests that Republican candidates are incentivized to employ more extremist rhetoric in the primary to rally the base and attain the nomination. Then, in order to remain competitive in the general election, Republican candidates are incentivized to temper this rhetoric, leading to the moderation observed in my empirical results. This lends credence to the view that political primaries perpetuate political polarization by rewarding more extreme candidates. Moreover, this asymmetry in partisanship of the bases may allow Democratic candidates

to adopt a relatively more moderate stance in the primary. In turn, the large gap in extremity between the average Democrat and Republican in the primary enables the Democratic candidates to remain constant throughout the electoral cycle. This account is consistent with a model of moderation that incorporates costs to ideological movement.

Although the analyses show no clear point in time where messaging changes, this does not rule out strategic moderation. For instance, many primaries are uncompetitive and essentially decided months before the actual election. Thus, candidates may choose to slowly alter their rhetoric, beginning this change during the primary before general election voters are conscious of it. As my data only captures a single election year, I am not able to fully separate strategic moderation from other factors that might have affected rhetoric. Nonetheless, the current results are clearly consistent with the moderation hypothesis. Future work should address whether the findings generalize to other election cycles and further identify the role of strategic considerations for the shift in messaging.

The remainder of the paper is organized as follows. Section 2 situates the paper within the existing literature and discusses competing theoretical hypotheses for why a political candidate may or may not moderate over the course of the election. Section 3 presents a theoretical model and frames my hypotheses. Section 4 details my methodologies and Section 5 describes my data sources and provides descriptive statistics about my samples. Section 6 validates the obtained predictions. Section 7 details my empirical specifications and 8 discusses the results. Section 9 concludes the paper.

## 2. Relevant Literature

Conventional political punditry suggests that to capture their party's nomination, candidates must run to the extremes in the primary to establish themselves as strong partisans and then, upon doing so, move to the middle to win over independent and more centrist general election voters. Nonetheless, conventional wisdom also suggests that candidates caught doing so are dogged by accusations of "flip-flopping," from which it may be difficult to recover.

Do either of these truisms have theoretical or empirical grounding? I proceed by considering theoretical results in the literature, followed by a discussion of existing empirical work regarding

the post-primary moderation hypothesis. Based on this consideration of the extant literature, I then situate the contribution of this paper and its methodology. In the subsequent section I present a theoretical model of moderation to elucidate my hypotheses.

#### A. *Why might candidates moderate?*

One primary approach for studying ideological positioning and divergence uses the Hotelling-Downs model (Hotelling, 1929; Downs, 1957). The central result of the model is the median voter theorem: both candidates maximize their chance of winning the election by converging to the position of the median voter. More concretely, consider a unidimensional policy space along the real number line, and for simplicity, let this be given by  $X = [0, 1]$  and further suppose there are two parties  $L$  and  $R$ . These voters are scattered along this interval according to a continuous density function  $f$  and median  $x^m$  such that their position on the interval represents their optimal bliss point. Voters possess single-peaked preferences with a maximum at their bliss point and their utility decreases monotonically with the distance from this point. In this simple Downsian model where parties only care about winning, parties will only have one equilibrium strategy, staking out positions at the median voter, i.e.  $L = R = x^m$ . In reality, the two major American political parties are significantly divergent in ideology and policy platforms, though there is evidence to suggest that the American two-party system results in more politically moderate candidates who are more concerned with independent voters than are candidates in multiparty systems (Burden, 2001).

The Hotelling-Downs model can also be adapted to account for the effects of a primary race preceding the general election. The simplest application of the Downsian model to this two-stage example merely adds an independent preliminary “primary” stage to the model and assumes voters are myopic. That is, any given candidate must first compete in the primary to secure the party’s nomination, and upon winning, then competes in the general electorate. Voters are not forward looking and simply choose the candidate closest to their ideal point in each period.

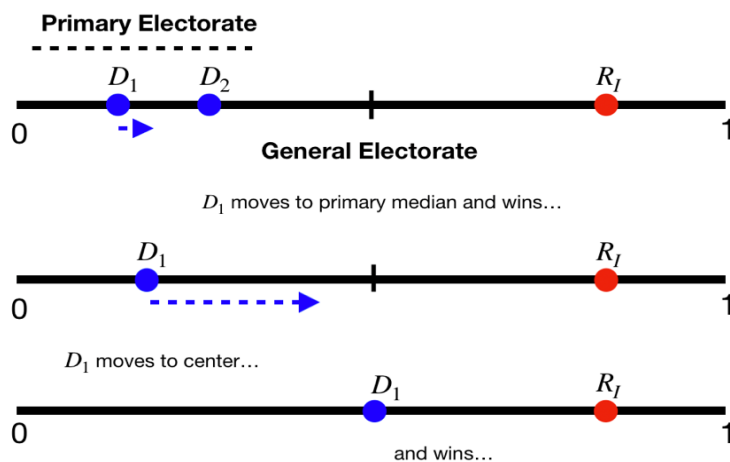
Cox (1990) identifies two sets of incentives that help to explain the candidate’s ideological positioning in each stage. Centrifugal forces push candidates toward the extremes of their party while centripetal forces pull candidates toward the center. In order to compete in the general election, candidates must successfully win over their primary voters. As primaries are dominated by parti-



san actors with relatively stronger ideological priors, centrifugal forces are paramount as candidates seek donations and support from the extreme end of the policy distribution. In the general election, the candidate must win over more independent voters and so the candidate is pulled toward the center. Hence, centripetal forces begin to dominate in the general.

In accord with Cox's identified dynamics, the equilibrium strategy for the candidates in this simple Downsian model is to converge to the median primary voter  $x_p^m$  and then if the candidate wins, shift toward the median general voter  $x_g^m$ . Under the assumption that party primary electorates are dominated by partisans, candidates should moderate between the primary and the general election, as illustrated in Figure 1.

FIGURE 1. MOVEMENT IN A DOWNSIAN CONTEXT WITH MYOPIC VOTERS



*Notes:* This figure illustrates the movement dynamics discussed in a Downsian model with myopic voters. In order to win the primary election, the candidate first moves to the median primary voter. Upon attaining the nomination, the candidate then moves to the median general voter. Here, voters are not forward-looking, but simply vote for the candidate closest to their ideal point at both stages of the model.

Importantly, a simple two-stage model such as this cannot account for potential interactive effects between the primary and general election stages. That is, in such a model, since voters are myopic and there are no barriers to ideological movement, the primary doesn't matter for the eventual general election outcome. This dynamic is quite unrealistic; for example, a politician's primary messaging may limit her messaging options for the general election because voters may punish her for having conflicting ideological messages, or alternatively, voters may not believe that a candidate's extreme signalling in the primary is genuine. To account for such effects, richer models

must be considered.

Hummel (2010) and Agranov (2016) both develop formal theories of two-stage elections that include possible flip-flopping costs. In particular, Hummel finds that the magnitude of moderation is determined by the associated costs of flip-flopping and that candidates always adopt divergent policies – that is, they do not fully converge to the median voter. Agranov shows the existence of an equilibrium where the post-primary moderation hypothesis holds, and primary winning probabilities are inversely related with winning probabilities in the general. Importantly, Agranov finds that if the competitiveness of the general election is sufficiently high, the primary candidates have no incentive to appear moderate. That is, all candidates will signal in the primary that they are an extreme type. Similarly, Hummel finds that in uncompetitive elections the upsides of any moderation strategy are dominated by the risks of flip-flopping and appearing dishonest. Thus, both of these models illustrate that candidates moderate over the course of the election even though voters are forward-looking and expect this moderation.

### *B. Why might candidates not moderate?*

Although the Downsian model predicts that candidates converge to the median voter, a robust empirical literature documents that in reality American politics is fairly polarized (Ansolabehere, Snyder Jr and Stewart III, 2001; Poole and Rosenthal, 2017). Returning to the one-stage model, there are multiple potential explanations for ideological divergence within each of the two parties. First, median voter predictions are difficult to maintain in multi-dimensional policy spaces (Plott, 1967). Assuming that voters are distributed along just three distinct policy spaces (e.g. economic issues, social issues, and foreign policy) renders a median voter equilibrium incredibly rare.

Indeed, the existence of this equilibrium only holds if the multidimensional space can be projected on a unidimensional space, ordering voters by type – an assumption that seems far-fetched given the heterogeneity of voter preferences across issues. Another potential explanation for the observed divergence from the median voter theorem in general elections is that parties are also ideologically motivated (Calvert, 1985; Wittman, 1977). For a brief example, consider an ideological space along  $[0, 1]$  with two parties  $L$  and  $R$ . Parties are policy-motivated according to the payoffs  $-\tau^2$  and  $-(1 - \tau)^2$  for a passed policy  $\tau$ . Parties do not know the location of the median voter  $x^m$ , and voters select the candidate closest to their ideal point. In this account, the unique solution exists

when party  $L$  chooses  $1/4$  and party  $R$  chooses  $3/4$  (assuming  $R > L$ ). Thus, when parties are also ideologically motivated, the equilibrium captures divergence. Although this result is from a one-period election cycle, it motivates the consideration of candidate preferences which may lead to ideological divergence and offers intuition for why candidates might not moderate.

Moreover, as noted by [Burden \(2004\)](#), candidates may choose not to adhere to the post-primary moderation hypothesis for a variety of other reasons: i) as mentioned above candidates may be ideologically motivated and are unwilling to compromise these convictions, ii) voters may actually punish candidates for flip-flopping on important issues, and iii) for incumbent candidates, they may have reputations that they need to protect or that may be difficult to transform. Similarly, [Tomz and Van Houweling \(2009, 2014\)](#) find empirical support that voters associate flip-flopping with negatively valenced trait characteristics, including dishonesty. Risk-averse voters may prefer, therefore, a candidate who commits to a policy, rather than someone who alternates between multiple policies. Many recent presidential candidates, including John Kerry, Hillary Clinton, and Mitt Romney, have all been dogged by accusations of flip-flopping on policy issues.<sup>3</sup>

Additionally, some candidates may run for ideological concerns, perhaps to get an issue on the table for discussion. Such candidates have little incentive to moderate, as winning the election is not their primary concern. Furthermore, in contexts where the race is not competitive and lacks intensity – due to strong incumbency effects or unpersuadable voters – there is little incentive for candidates to adopt extreme positions and then moderate, instead preferring to keep their position hidden ([Agranov, 2016](#)).

### *C. Empirical Literature*

As mentioned above, the political science literature contains limited empirical evidence on the existence of moderation among congressional candidates. This absence is primarily driven by the difficulty in obtaining well-defined quantitative measures of political ideology for congressional candidates, who do not have a roll-call history. Thus, the majority of research in this field has focused on incumbent candidates.

---

<sup>3</sup>These accusations of “flip-flopping” are treated seriously by the candidates and receive significant media attention. For example, Mitt Romney [addressed](#) his accusations at the first Town Hall of the 2012 presidential primary, though it continued to follow him (see this [Rolling Stone article](#) for a compilation). For coverage of “flip-flopping” for Clinton, see write-ups in the [Washington Post](#), [NPR](#), and [Politico](#). Additionally, see [Croco \(2016\)](#). Moreover, [Jones \(2011\)](#) writes “Members of Congress believe that their electoral fortunes depend significantly on the positions they take.”

Burden (2001) is an important work in this area, focusing on the extent of ideological moderation among incumbents. In particular, Burden calculated DW-Nominate estimates for each incumbent for both the primary and the general period, based on their simultaneous roll-call voting behavior and found slight changes in legislator positions toward moderation. These shifts were greater among Republicans. Nevertheless, this ideological estimation methodology is clearly limited in scope, not simply because it is restricted to incumbents but also as it presupposes an electorate that is conscious of and reactive to everyday congressional voting patterns. While high-profile bills may attract media attention and elicit voter accountability (Ansolabehere and Jones, 2010), minor shifts in roll-call voting concurrent with the election cycle are likely not perceived by voters.<sup>4</sup>

Following up on this work, Burden (2004) developed a mail survey, sent out to every major party candidate, shortly before the 2000 general election, which consisted of a single task: to place themselves on a left-right ideological scale. The survey results indicated significant divergence among Democrats and Republicans, incumbents and challengers alike, though candidate positions in competitive elections were more convergent. Burden interpreted this as an indication that candidates do not moderate over the course of the election cycle, though these findings do not preclude the fact that candidates could have staked out even more extreme positions in the primary. Moreover, using self-placement poses problems, not just because one factor impacting the score was which staff member completed the survey, but also because the ideological scale is not standardized. For example, people may be inclined to report themselves as more moderate than they truly are. In addition, the incentives to respond truthfully are not clear, candidates may perceive benefit from pandering here, too.

More recently, Acree et al. (2020) investigated the extent of ideological moderation in the presidential elections in 2008 and 2012 using a two-stage text analysis model. Since the authors only focused on the presidential race, they were able to take advantage of a rich archive of speeches. The authors model the problem as a classification task, specifying different ideological classes (for example, on the left: “center-left”, “religious left”, “progressive”, and “far left”), within which they attempt to classify the candidates during different time periods. They construct their baseline

---

<sup>4</sup>For example, Guisinger (2009) shows that voters have little knowledge of their representative’s positions on trade policy. In addition, survey materials regularly show that voters are uninformed about their representatives. For example, a 2017 survey found that only 56% of participants could provide the correct party affiliation for the representative in their congressional district.

model using a large corpus of political texts, which they labelled according to the classes above. The authors' results identify the candidates as more ideologically extreme during the primary season than in the general. However, the measure developed by [Acree et al. \(2020\)](#) yields somewhat implausible results. For example, Obama is estimated to be employing conservative rhetoric more regularly in the 2008 general than his Republican opponents in the 2008 and 2012 general elections.

In the arena of congressional elections, such a robust archive of speeches does not exist. As a result, many studies have also relied on donation data ([Bonica, 2013, 2014](#); [McCarty, Poole and Rosenthal, 2016](#)) from the FEC, estimating the political leaning of donors and then using the weighted average of donors to a candidate to calculate their ideology. While these donation-based estimates largely coincide with DW-Nominate ratings, they are not well suited for comparison between the primary and the general. First, strategic donations are a serious concern. That is, donors may donate to get access to a politician or based on electability concerns, rather than donate to further ideological interests. In fact, [Hall and Snyder \(2015\)](#) conclude that donors act even more strategically than voters, “wasting” very few donations on candidates outside of the top two in primaries. As it is not possible to distinguish between ideological and strategic donations, this issue has the potential to bias ideology estimates. Second, it stands to reason that donors are on average substantially wealthier than non-donors. This likely biases results in the direction of elites: for example, donation history will make campaigns look less populist than their rhetoric may suggest.

Finally, differences in the composition of voters in the general and primary significantly bias the result. For example, in the general, partisans will likely coalesce around their party nominee regardless of ideology and independents – that sat out the primary – will choose a side. To provide some intuition for this, consider the example of a primary with two Democratic candidates,  $D_1$  and  $D_2$  positioned at  $x_1$  and  $x_2$  respectively, a Republican candidate  $R$  at  $x_R$ , and three primary donors/voters, located at  $y_1, y_2$ , and  $y_3$ . Let  $x_1 = y_1 < x_2 = y_2 = y_3 < x_R$ . Let voters choose the candidate that is closest to them, and thus,  $y_1$  prefers  $x_1$  and  $y_2$  and  $y_3$  prefer  $x_2$ , and so candidate  $x_2$  wins the primary. Assuming that the voters donate to their preferred candidate, this methodology yields an observed ideological position of  $y_2$  for  $x_2$ . For the general election, suppose  $x_2$  does not change their position for the general election. The voter  $y_1$  still prefers  $x_2$  to  $x_R$  and donates accordingly. The observed ideological position of  $x_2$  is now observed as  $\frac{1}{3}(2y_2 + y_1)$

even though the  $x_2$ 's true position remained unchanged. A similar example can be constructed for independents, as independents likely do not participate in the primary election but then donate to a candidate in the general election, artificially biasing the candidate's observed position closer to the independents' bliss point (even if the candidate's ideology actually remains unchanged). Hence, donation data do not provide a suitable metric for comparison from the primary to the general election.

#### *D. Contribution*

This paper focuses on candidates' speech in social media, exploiting its rapid proliferation as a tool for political communication. This focus on social media interactions as opposed to congressional speeches or policy platforms is a key contribution of this study. While there would also be merit in analyzing the policy platforms of candidates – as platforms signal a candidate's priorities and are considerably different across the ideological spectrum – the proposed focus on social media has significant advantages. Most importantly, focusing on tweets provides an opportunity to analyze the ideological rhetoric of candidates, which is more flexible than the policy positions outlined in policy platforms. For instance, a candidate can discuss more extreme policies that play well in the primary to rally support among partisans and attain the nomination (“red-meat issues”), and then in the general alter their speech using more balanced rhetoric and bipartisan appeals to sound more moderate.

Additionally, candidates can adapt the reasoning they use to justify their positions, essentially repackaging the ideas in wrapping more suitable for the electorate. [Acree et al. \(2020\)](#) argue that such rhetorical shifts are less likely to trigger “flip-flopping” accusations than actual policy shifts. Moreover, candidates may stress different parts of their agenda depending on the audience. For example, a Republican primary candidate may propose to build a wall along the Mexican border in the primary, but upon advancing to the general election, pivot to emphasizing plans to cut middle-class taxes as a more moderate appeal to independent voters. Finally, candidates may even be able to exploit informational asymmetries of general election voters to make their policies appear more centrist than they really are ([Iyengar and Simon, 2000](#)).

In addition, Twitter data are advantageous for several other reasons. First, social media presents the opportunity to observe candidates in real dialogue with their potential constituents. That

is, policy platforms and speeches may rely on a nuanced understanding of policy and may only be studied carefully by policy wonks and highly educated or interested constituents. In contrast, social media provides a channel for candidates to directly engage a broader constituency in a more accessible manner. Second, focusing on tweets provides a more continuous estimation of the candidate's evolving ideological positioning than would an analysis of policy platforms. If a candidate changes their policy platform and not just their rhetoric, they will likely remove all references to the prior platform. In contrast, given the high volume of tweets, candidates are unlikely to delete individual tweets from months prior.

Thus, my focus on candidate rhetoric on Twitter allows me to address the lack of empirical evidence on ideological movement over the election cycle. In particular, it enables me to offer a generalized methodological approach for measuring ideology over time, which – unlike most work in the literature – can be extended universally to political candidates. In turn, I am able to contribute one of the first empirical analyses on the evolution of congressional candidates' ideological rhetoric in order to provide quantitative evidence on the post-primary moderation hypothesis.

It is important to acknowledge that the population of Twitter users is different than the voting population. Indeed, Twitter users are younger, wealthier, and more educated than the average American adult (Pew, 2019). As a result, it is a reasonable concern that candidates might behave or invoke different rhetoric on the platform than off the platform. While I cannot directly address this, [Yaqub et al. \(2017\)](#) show that political Twitter discussions are correlated with both real world events and public events pertaining to the elections. Moreover, I argue that Twitter has sufficient influence and importance to the political environment that even if rhetorical shifts only occurred on the platform, it would still be notable and represent a substantive movement. For example, [Buccoliero et al. \(2020\)](#) assert that Twitter “drove rather than merely followed the developments of the [2016] presidential election.”<sup>5</sup> These studies are part of a substantial political science literature on the usage of Twitter – and social media more broadly – as an important tool for political communication (for an extensive overview of this literature, see [Jungherr, 2016](#); [Kreiss and McGregor, 2018](#)).

---

<sup>5</sup>Writing for *The Hill*, [McCabe \(2015\)](#) terms the 2016 presidential election “the social media election” due to the relevance of Twitter and other social media platforms to the campaigns.

### 3. Model

In this section, I describe the basic features of the [Agranov \(2016\)](#) incomplete information model of two-stage elections to provide intuition underlying the post-primary moderation hypothesis when both candidates and voters are more sophisticated than the simple Downsian model.

The model is structured with two Democratic candidates  $c^A$  and  $c^B$ , a continuum of voters, and an incumbent Republican candidate with type  $R$ . The two candidates are assigned either a liberal ( $L$ ) or moderate ( $M$ ) type with equal probability. The candidates' locations satisfy the inequality  $0 < L < M < R < 1$ . Each voter  $j$  has a policy ideal point  $z_j$  along the interval  $[0, 1]$  such that those falling between  $[0, \bar{z}]$  are Democrats. The median primary voter  $d \in [0, \bar{z}]$  and general voter  $m \in [0, 1]$  are drawn from cumulative distribution functions but are not known to the candidates. However, it is assumed that the median general voter is an independent voter, that is,  $m > \bar{z}$ .

In the primary, each candidate  $k$  chooses an effort level  $e_1^k \in [0, 1]$ , which generates a signal  $s_1^k \in \{\lambda, \mu\}$  to the voters about  $k$ 's type, where  $\lambda$  is a liberal and  $\mu$  a moderate signal. This effort level can be thought of as the campaigning cost borne by the candidate to persuade the voters of their true type. In particular, the probability of generating the opposite type signal is increasing in the effort level; when the effort level is zero, the generated signal is equal to the candidate's true type. The success of candidates distorting their signal is determined by the prominence of the election  $n_i > 0$ ; thus, in more important elections it is more costly for the candidate to manipulate their signal. Note that the general is always more prominent than the primary ( $n_2 > n_1$ ).

In response to these signals, voters cast their vote for their preferred candidate, and the winner then proceeds to the general. In the general, the winning candidate  $w$  chooses an effort level  $e_1^w \in [0, 1]$  for the campaign and a signal  $s_1^w \in \{\lambda, \mu\}$  is generated to all voters. The type of the incumbent is known. Once again, voters cast their ballots and the winner implements a policy according to their type. At neither stage are voters able to abstain.

Candidates receive utility for winning the general and disutility for expending effort according to

$$\Pi^k(e_1^k, e_2^k) = 1_{\{k \text{ won general}\}} - e_1^k - e_2^k.$$

Voters receive utility according to the distance of the passed policy from their ideal point; that is,



for voter  $j$  with ideal point  $z_j \in [0, 1]$ , their utility is defined as

$$u(z_j, p) = -(z_j - p)^2$$

for policy  $p$ . Further, it is assumed that the median primary voter prefers the liberal type to the moderate type, but all primary voters prefer either Democratic candidate to the incumbent in the general.

Agranov (2016) shows that this model has a “pandering equilibrium” where candidates pander according to the preferred type in each stage of the election. During the primary liberal candidates exert no effort while moderate candidates expend effort to imitate the liberal type. As for the primary winner, a candidate generating a liberal signal always wins against a candidate generating a moderate signal. In the general, if the winning primary candidate generated a liberal signal, they will exert effort according to their true type: if moderate, they will do nothing; else they will attempt to mimic a moderate signal. Similarly, in the case that the primary winner generated a moderate signal, they will exert no effort in the general.

Given these dynamics, increasing the prominence of the primary election decreases the probability that a moderate candidate will win, as the effort required to generate a liberal signal becomes too costly. Moreover, it discourages liberal types from expending effort in the general, as voters are less likely to believe the challenger is moderate after the prominent primary. Analogously, in settings where the general election is of high prominence, moderate challengers have a higher chance of winning in the general, and thus will also expend more effort in the primary to disguise their signal.

Since congressional primaries are not of particularly high salience, especially from the perspective of independent general election voters, the pandering equilibrium suggests that candidates will attempt to appear more extreme in the primary to attain the nomination. Upon winning, candidates will then attempt to generate a moderate signal for the general. This dynamic is homologous to the behavior predicted by the post-primary moderation hypothesis.

Moreover, it is natural to consider these comparative statics in the context of candidate- and race-specific attributes. For instance, competitive generals can be modeled as races of especially high second-stage prominence. Consequently, I expect moderate candidates to expend considerably

more effort to distort their true signal in the primary for races with competitive general elections, thereby accentuating the moderation effect over the course of the cycle. In addition, by allowing voters to be forward-looking and to update their beliefs, this model effectively incorporates flip-flopping costs, as candidates that successfully distort their type in the primary will also have counter-productively shifted the beliefs of general election voters. That is, for a candidate of moderate type, attempting to generate a liberal signal in the primary will simultaneously increase the probability that general election voters place on the candidate's true type being liberal. Thus – abusing notation slightly, letting the incumbent occupy a candidate role  $c^A$  and the opposite party primary winner fixed as the “Republican incumbent” role from the original model – I extend the model for the situation of incumbent candidate facing a primary challenger where the incumbent candidate bears an asymmetrically higher prominence score than their challenger. Hence, the incumbent is more constrained in distorting their signal.

To synthesize these intuitions, the model provides a framework for understanding the post-primary moderation hypothesis: that is, that candidates will appear more extreme in the primary and then moderate for the general election. Using the model's notion of prominence, I extend this to argue that the effect should be especially pronounced in competitive general elections. Finally, I consider an asymmetric level of prominence as indicative of greater flip-flopping costs borne by incumbents, which should consequently constrain incumbents from moderating as much as non-incumbents. As a result, I form the following hypotheses:

- 1) Congressional candidates should moderate from the primary to the general election.
- 2) The extent of moderation among congressional candidates in races with competitive general elections should exceed that of those in uncompetitive races.
- 3) Moderation among incumbent candidates should be less than among non-incumbents.

## 4. Methods

In order to translate text data into estimates of ideological extremity, I employ three different approaches. First, I utilize a bag-of-words approach to the text data, removing the grammatical structure, and converting it into frequencies of bigrams. From this, I select the most partisan bigrams as features using a Chi-Squared test and specify a Multinomial Inverse Regression to build

a model of speech that predicts partisanship based on the occurrences of these bigrams. Second, I take advantage of the Moral Foundations Framework dictionary, an externally validated and verified set of keywords representing different moral foundations, to obtain the frequencies with which candidates invoke moral values, such as fairness or authority. Given a candidates’ basket of frequencies, I can construct a measure of the relative importance of universalist values, which has been shown to be highly correlated to ideology. Finally, I specify a natural language of model using a pre-trained RoBERTa Transformer model, which I fine-tune on the task of ideological prediction. In this case, these predictions are based on a minimally-cleaned text data that retains all of its original context, sentence, and grammatical structure.

Throughout this section, I will refer to my baseline sample and my candidate sample. The baseline sample corresponds to the set of Representatives from the 116th Congress and the candidate sample to the set of candidates running for congressional office in 2020. The collection and construction of these samples is described in greater detail in Section 5.

### A. *Multinomial Inverse Regression (MNIR) Approach*

#### FEATURE SELECTION

Since this method takes a bag-of-words approach, the text must be substantially cleaned for text analysis algorithms that do not consider grammatical and sentence structure. Thus, I proceed with standard text-analysis pre-processing steps (Gentzkow, Shapiro and Taddy, 2019). This includes removing hyphens and apostrophes and replacing all other punctuation with spaces as well as extremely common words from a list of stopwords (such as “the”, “a”, etc.).<sup>6</sup> I then apply the Porter stemming algorithm to reduce words of a common stem (Porter, 1980). The text is then converted into overlapping two-word phrases (bigrams).

For example, the sentence “I do not approve of death taxes.” would yield the bigram representations “approve death” and “death tax”, as the words “I”, “do”, “not”, and “of” would all be contained in the set of stop words while “taxes” would be coded as its stem “tax”. One limitation of this text processing procedure is that the contradictory sentence “I approve of death taxes” would yield the exact same bigram representations. Nonetheless, so long as the affirmative sentence

---

<sup>6</sup>This list can be obtained from the `nltk` Python module.

appears systematically more frequently than the negative framing or vice versa and this discrepancy falls along partisan lines, there will still be an important statistical signal for the model to learn. Moreover, there is evidence that partisans use ideologically-loaded terminology – for example, Democrats use the phrase “estate tax” whereas Republicans say “death tax” (Gentzkow and Shapiro, 2010).

After completing these text processing steps, I obtain a list of bigrams and their corresponding use frequencies for each congressperson in the baseline sample and for each candidate-time observation in the main sample. In order to ascertain a measure of the ideological valence of the candidates, I will compare the relative bigram frequencies among the candidates to the frequencies in the baseline sample to identify the extent to which the candidate’s language is more characteristic of a Republican or Democrat. While this could be computed over the entire set of bigrams in the baseline sample, this would be extremely computationally costly and would likely include many bigrams with negligible ideological signal.<sup>7</sup> Instead, I complete a feature selection step to identify the 10,000 most partisan bigrams in the baseline data according to a Chi-squared test. In particular, I select bigrams according to the following procedure.

Define  $c_{jp}$  to be the frequency of bigram  $j$  and  $c_{\sim jp}$  the the total frequency of all bigrams other than  $j$  for party  $p$  in my baseline sample where  $p \in \{d, r\}$ . Then the Chi-squared statistic  $\chi_j^2$  is defined as

$$\chi_j^2 = \frac{(c_{jr}c_{\sim jd} - c_{jd}c_{\sim jr})^2}{(c_{jr} + c_{jd})(c_{jr} + c_{\sim jr})(c_{jd} + c_{\sim jd})(c_{\sim jr} + c_{\sim jd})}$$

Assuming that the bigram frequencies  $c_{jp}$  are drawn from multinomial distributions  $MN_{jp}$ , the statistic  $\chi_j^2$  can be interpreted as the test statistic of the null hypothesis that the likelihood of using bigram  $j$  is identical between Republicans and Democrats (Gentzkow and Shapiro, 2010). Thus, large values of  $\chi_j^2$  suggest that the usage of phrase  $j$  is highly partisan, whereas small values of the test statistic suggest that  $j$ 's usage is not diagnostic of party membership. For example, the bigrams with the largest  $\chi^2$  values are “trump\_administr” and “gun\_violenc” with 12,149 and 8646 mentions by Democrats and only 1,034 and 165 by Republicans, respectively, while the bigram with lowest  $\chi^2$  is “tough time” with 96 mentions by Democrats and 55 by Republicans. This makes intuitive sense, as it would not make sense to infer partisanship from a non-political term like

---

<sup>7</sup>Overall, there are over 3 million bigrams in the baseline sample. Of these, there are over 70,000 that have been mentioned at least 20 times.

“tough\_time”. By selecting the top 10,000 bigrams according to the Chi-squared statistic, I am able to drop these nonpartisan phrases. The most partisan phrases are displayed in Table 4 in the validation analysis (Section 6).

For the feature selection, I first drop all bigrams that do not occur more than 20 times ( $c_{jr} + c_{jd} < 20$ ). This is to ensure that the selected phrases are not specific to an individual speaker and are sufficiently common to have a reasonable likelihood of occurring in the candidates’ tweets as well. Phrases that don’t meet this criteria would not be diagnostic and thus I remove them. Of the remaining bigrams, I select the 10,000 with the greatest values of  $\chi_j^2$ . Given the relatively small word content of individual tweets (limited to 240 characters), a large number of bigrams were chosen to increase the number of matched phrases among the candidates’ tweets and thereby lower the variance of the predictions.<sup>8</sup> However, the exact cutoffs were chosen arbitrarily.

#### MULTINOMIAL INVERSE REGRESSION

I proceed by taking advantage of the observed word count matrix from the baseline sample to build a model of speech to predict the partisanship of each candidate in both the primary and the general. In particular, I choose to model partisanship using the Multinomial Inverse Regression (MNIR) from Taddy (2013). MNIR has many desirable properties. Foremost, it is tractable and has been successfully used previously to extract meaning from text, specifically in the context of ideological prediction (Taddy, 2013, 2015; Gentzkow, Kelly and Taddy, 2019; Gentzkow, Shapiro and Taddy, 2019). Moreover, it allows for dimension reduction, significantly increasing computational speed, while maintaining estimates of sentiment. Finally, the multinomial model is a natural model for speech, as it supposes that people choose phrases and combinations of words instead of assuming the independence of individual words. However, it is also important to acknowledge the necessary assumptions of this model, particularly that each candidate’s tweets are independent of all other candidates.

Let  $\mathcal{J}$  be the set of bigrams obtained from the feature selection step. Then, for each speaker  $i$ , I observe the vector of counts  $c_{is}$  of selected bigrams where  $s$  refers to the session of Congress. As my sample only considers one such session, the index  $s$  will be omitted for brevity. I define the

---

<sup>8</sup>For example, as will be discussed later on in the validation section, with this set of 10,000 bigrams obtained from the baseline sample, the median candidate-month observation among Democrats has 56 overall counts and for Republicans only 36.

total amount of speech for each speaker to be

$$m_i := \sum_{j \in \mathcal{J}} c_{ij}.$$

Then, I assume that the phrase counts vector for each speaker is distributed according to the multinomial distribution:

$$c_i \sim \text{MN}(m_i, q_i(X_i))$$

where  $X_i$  is a vector of  $k$  speaker covariates and  $q_i(\cdot)$  is the probability distribution over the bigrams. The probability of speaking phrase  $j$  is given as

$$q_{ij}(X_i) = \frac{\exp\left(\alpha_j + \sum_{l=1}^k \varphi_{jl} X_{il}\right)}{\sum_{j \in \mathcal{J}} \exp\left(\alpha_j + \sum_{l=1}^k \varphi_{jl} X_{il}\right)}$$

where  $\varphi_{jl}$  represents the coefficient on variable  $l$ . Here,  $\varphi_{jl}$  can be interpreted as the effect of the speaker characteristic  $l$  on the propensity to use the phrase  $j$ . The scalar parameter  $\alpha_j$  captures the baseline popularity of the phrase  $j$ . In my main specification,  $X_i$  includes an indicator for the speaker's party, the DW-Nominate 1 score, the DW-Nominate 2 score, an interaction between party and each DW-Nominate score, and the speaker's age.

Estimation of the multinomial regression above is computationally difficult. In order to ease computation, I approximate the likelihood of the counts with the likelihood of a Poisson model (as in [Gentzkow, Shapiro and Taddy, 2019](#)) and make use of a penalized objective function with an  $L_1$  penalty in order to impose sparsity on the loadings. The usage of the sharp  $L_1$  penalty implies that many of the loadings on the bigrams can be expected to be zero, which should reduce overfitting.

As detailed in [Taddy \(2013\)](#) in order to obtain a lower dimensionality, I calculate a sufficient reduction (SR) score as given by

$$Z_i = \varphi \frac{c_i}{m_i}$$

where, for all  $l$ ,  $X_{il} \perp c_i, m_i \mid Z_i$  and  $\varphi$  is the coefficient matrix at the bigram level using mentions across all speakers. The resulting score  $Z_i$  is of dimension  $k$ , that is a SR score is calculated for each covariate included. This reduction is computationally efficient, as it allows us to directly model the party affiliation of text against the SR projection values.

Now that I have constructed the multinomial distribution and reduced the dimension of the speech vectors, I can use the observed speech counts for the representatives to project their position on the DW-Nominate scale. To do so, I estimate the simple forward linear regression model

$$y_i = \beta_0 + \sum_{l=1}^k \beta_l Z_{il}$$

on the baseline sample where  $y_i$  represents the DW-Nominate Score. I fit this relationship using a variety of methods, including linear regression, regression forest, and gradient boosting. The latter two methods have better out-of-sample performance than linear regression, as measured on a validation set of sitting Senators. This is likely as both include regularization terms to prevent overfitting.

With the forward regression, I obtain a mapping from speech counts to DW-Nominate 1 scores. Hence, for each candidate-month observation I can calculate the dimension-reduced scores  $Z_{it}$  from their observed counts and use the coefficients from the forward regression above to obtain a prediction of their ideology relative to the baseline sample. These fitted DW-Nominate scores  $y_{it}$  will make up my outcome variable in my regression specifications.

### *B. Moral Foundations (MFD) Approach*

Rather than relying on automated feature selection steps, this approach is theoretically-driven, using a set of pre-verified dictionaries to obtain estimates of the relative frequency that candidates invoke universalist or communal values over the course of the election cycle. This dichotomy between universalism and communalism has been shown to be closely related to ideology ([Haidt and Graham, 2007](#); [Graham, Haidt and Nosek, 2009](#); [Enke, 2020](#)). For example, [Graham, Haidt and Nosek \(2009\)](#) show that political conservatives are much more likely to hold communal values than liberals. Furthermore, analyzing recent presidential elections, [Enke \(2020\)](#) demonstrates that the Democratic party candidates invoke universalist rhetoric significantly more often than the Republican party candidates. Additionally, using a survey design, Enke shows that on the voter level a standard deviation increase in the value of universalism is associated with a ten percentage point decrease in the probability of voting for Trump.

The main advantage of this approach is that it provides a dictionary of keywords that have

been theoretically-validated for the purpose of measuring moral values, which in turn correspond to political ideology. To determine these moral values, I use the Moral Foundations Theory (MFT) framework of morality (Haidt and Joseph, 2004; Graham, Haidt and Nosek, 2009; Graham et al., 2011). This framework sets out that moral concerns can be partitioned into five foundations. These are

- 1) Care/Harm: the extent to which people care for the weak and attempt to keep others from harm.
- 2) Fairness/Reciprocity: the importance of ideas relating to equality, justice, rights, and autonomy.
- 3) In-group/Loyalty: people's emphasis on being loyal to the "in-group" (family, country), patriotism, and the moral relevance of betrayal.
- 4) Authority/Respect: the importance of respect for authority, tradition, and societal order.
- 5) Purity/Sanctity: the importance of ideas related to purity, disgust, and traditional religious attitudes.

as described in Enke (2020). Each of these dimensions has a virtue and a vice framing. For example, for Care/Harm, Care is the virtue framing and includes a word like "security" whereas Harm is the vice framing and includes a term like "suffering".

The first two dimensions, namely Care/Harm and Fairness/Reciprocity, correspond to universalist rhetoric while the next two dimensions (In-group/Loyalty and Authority/Respect) can be thought of as communal. This distinction lies in the fact that the values from the first two foundations are not conditional based on certain groups or relationships, whereas the latter two are thought to be ethics of a certain group or community (Graham et al., 2011). In other words, the values in the universalist set are extended to all people, no matter their identity. Values of caring for the weak or ensuring fairness hold no matter the background of the target. In contrast, feelings of patriotism do not extend to all countries, rather they are specific to a person's country of origin or residence. Finally, because it does not have a clear relationship to universalist or communal rhetoric, the Purity/Sanctity dimension is not considered in this analysis.

The claim that moral values can be partitioned into these exact five foundations is still an area for



active research within the psychological literature. For example, the authors of MFT acknowledge that they are considering an additional foundation based on Liberty/Oppression, as studied in (Iyer et al., 2012).<sup>9</sup> Nonetheless, there is broad agreement in the psychological and philosophical literature for the distinction between universalist and communal moral values (for recent examples using this paradigm, see Luguri, Napier and Dovidio, 2012; Hofmann et al., 2014; Smith et al., 2014; Hannikainen, Miller and Cushman, 2017; Enke, 2020; Enke, Rodriguez-Padilla and Zimmermann, 2021).

To obtain estimates of moral rhetoric, I use the set of moral keywords created by Graham, Haidt and Nosek (2009).<sup>10</sup> This forms the Moral Foundations Dictionary (MFD). Each category contains about 10-20 keywords for the virtue and for the vice framing. I am then able to count the occurrences of these keywords in each candidates' speech over time. As the bulk of the text is discarded and only the keyword occurrences are stored, this approach does not require any text processing steps.

As my summary statistic, I use Enke's statistic of the relative importance of universalist versus communal values, adapted by changing the sign in order to match the dynamics of DW-Nominate where positive values are conservative:

$$\begin{aligned} \text{Relative importance of communal values} &= - [\text{Relative importance of universalist values}] \\ &= - [\text{Universalist values} - \text{Communal values}] \\ &= \text{In-group} + \text{Authority} - \text{Care} - \text{Fairness}. \end{aligned}$$

More specifically, I define the left hand side quantity  $u$  to be

$$u = \frac{f_{\text{Ingroup}} + f_{\text{Authority}} - f_{\text{Care}} - f_{\text{Fairness}}}{\text{Total number of non-stop words}},$$

where

$$f_i = \frac{1}{2} \left( \frac{1}{N_i^v} \sum_{z_i=1}^{N_i^v} n_{z_i} + \frac{1}{N_i^m} \sum_{z_i=1}^{N_i^m} n_{z_i} \right)$$

<sup>9</sup>This value is intended to capture "feelings of resentment toward those who dominate them and restrict their liberty." For more information, see <https://moralfoundations.org>.

<sup>10</sup>This dictionary is available at <https://moralfoundations.org/other-materials>.

and  $z_i$  represents keyword  $z$  for themes  $i$ ,  $n_{z_i}$  its frequency, and finally  $N_i^v$  to be the total number of vice words for theme  $i$  and  $N_i^m$  the number of virtue words.

Because of the small cardinality of the keyword set, I aggregate the occurrences at the candidate-month level. Using each candidate's ( $c$ ) tweets for each month ( $t$ ), I compute the frequency of their universalist rhetoric  $u_{c,t}$ . With the weighted mean  $\mu_u$  and standard deviation  $\sigma_u$  over these values, I compute the z-score

$$z_{c,t} = \frac{u_{c,t} - \mu_u}{\sigma_u},$$

where the weighting measure is given by the usage of the overall number of occurrences of keywords,

$$w = \sum_i \left[ \sum_{z_i=1}^{N_i^v} n_{z_i} + \sum_{z_i=1}^{N_i^m} n_{z_i} \right].$$

Importantly, since this method relies on simple frequencies of keywords, it does not need to be trained with the baseline sample. Consequently, it is not dependent on the assumption that the baseline sample of Tweets is generalizable to the congressional candidates' Tweets.

Finally, to match the range of the predictions from the MNIR section, I winsorize the  $z$ -scores at three standard deviations and scale them to be bounded between -1 and 1. The resulting values,  $y$ , will form my outcome variable for this methodology.

### C. Natural Language Model (RoBERTa) Approach

The two methods for assessing candidate ideology I have described thus far are selective by design. They each rely on only a subset of actual speech, selecting relevant text either statistically as in the bag-of-words approach, or by referencing an a-priori defined set of keywords in the case of the MFD-based measure. In contrast, the third approach I develop to measuring candidate ideology allows me to specify a full natural language model that takes into account context and grammatical structure. As a result, the text is only very lightly processed and sentences are preserved as they appeared.<sup>11</sup>

For this analysis, I use a deep learning Transformer model, RoBERTa, with multiple encoding

---

<sup>11</sup>To ensure that the inputs are well-formed, I remove any URL links and Twitter-specific content, such as "@" for handles and "#" for hashtags.

levels that uses self-attention, essentially allowing the model to “remember” and attend to previous words in the input stream. RoBERTa is an improved version of Bidirectional Encoder Representations from Transformers (BERT) (Liu et al., 2019), and is available through the Huggingfaces module.<sup>12</sup> These models are pre-trained on an extensive corpus of language and then made available for researchers to use on downstream tasks, such as ideological prediction. They provide state-of-the-art performance on almost all natural language processing (NLP) tasks.

In pretraining, BERT models take as input a pair of sentences from their source data, represented as sequences of tokens, which I denote as  $x = (x_1, \dots, x_N)$  and  $y = (y_1, \dots, y_M)$ . These are concatenated to form a single input sequence with special characters to delimit them. The models use a transformer architecture as depicted in Figure 2 (Vaswani et al., 2017). This architecture makes use of an encoder-decoder structure, where the input sequence is first mapped to some continuous internal representation (“encoder”), from which the model then generates the output tokens sequentially (“decoder”). The motivation behind this structure is perhaps most easily understood in the domain of machine translation where the model simultaneously learns to map the source language input to an underlying encoding vector space, and from this intermediate space to an output in the target language. However, this structure has been quite successful across a variety of natural language processing tasks, not simply limited to translation.

The model uses stacked self-attention, which consists of running multiple attention mechanisms in parallel and pooling the outputs. Each attention layer is fully connected, so every element in the input interacts with every other element in the input. The resulting attention scores can be interpreted as representing the importance of each element in the given input, as it captures what the model focuses on when computing the output. Thus, in combining these self-attention layers, the architecture is enabling the model to learn to “attend to” different aspects of the input stream with the most signal simultaneously. The exact attention mechanism utilized is referred to as a scaled dot-product attention, which combines the attention inputs through matrix multiplication, visualized in Figure 2. For a more detailed discussion of this mechanism, see Vaswani et al. (2017).

In natural language processing, machine learning models generally specify an objective to predict the next word given a context of surrounding words. This enables the models to learn rich

---

<sup>12</sup>See [https://huggingface.co/docs/transformers/model\\_doc/roberta](https://huggingface.co/docs/transformers/model_doc/roberta) for more information.

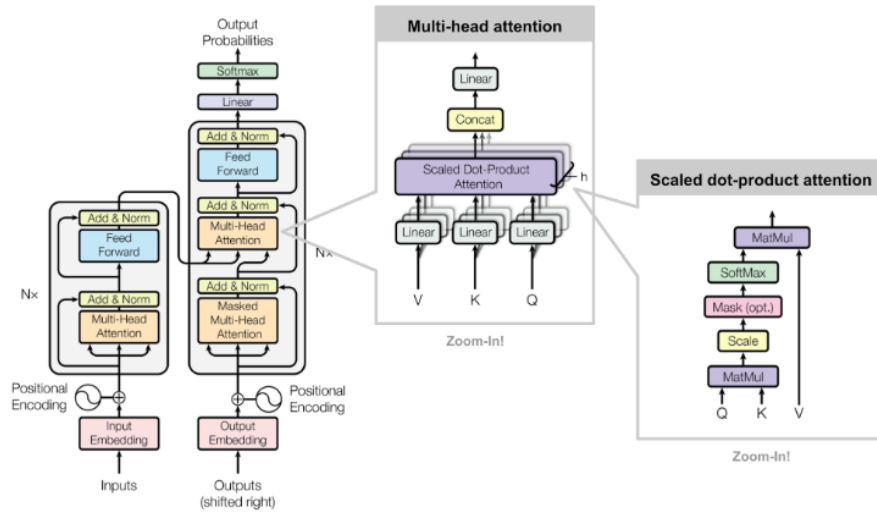


FIGURE 2. TRANSFORMER MODEL ARCHITECTURE

*Notes:* This figure presents the Transformer model architecture, as presented in Vaswani et al. (2017), illustrating the encoder-decoder structure of the model (left) with the multi-head attention and the attention mechanism magnified. For a more detailed discussion of this architecture, see the original paper.

dependencies between words and ultimately sentence structures (Mikolov et al., 2013). In contrast to traditional neural NLP methods, such as recurrent and long short-term memory neural networks, which process a sequence token-by-token and attempt to predict using either a left- or right-context, BERT models use a bidirectional context as part of its Masked Language Modeling (MLM) objective. As an example, consider the sentence “I do not approve of death taxes” with target word “approve”. The left-context of this model would be “I do not” and the right-context would be “of death taxes.” In addition, these contexts are obtained sequentially, thus words closer to the target word would have more weight. Meanwhile, the bidirectional context would mask the word “approve” and then process the input “I do not [MASK] of death taxes.” More generally,  $k$  percent of the words in input sequences to BERT are masked and then processed at once, thereby enabling the model to predict these masked words using the full context. This bidirectional context enables the model to exercise self-attention and removes the locality bias, where, for a given token, tokens that are closer are given more weight. In practice, 15 percent of the words are selected for the prediction, of which 80 percent are masked, 10 are replaced with a random word, and the remaining 10 with the actual word. The original BERT implementation uses static masking, performing masking only once during data preprocessing. This data was then duplicated ten times such that each sequence is masked in ten different ways in training (Devlin et al., 2018). The

implementation of RoBERTa alters this slightly, by replacing the static masking in BERT with dynamic masking, where the masking pattern is regenerated every time a sequence is fed to the model. Thus, the masked word is different across the instances of the training data, leading to more robust performance.

In addition, the initial BERT implementation also specifies a Next Sentence Prediction (NSP) objective, which is used to learn relationships between sentences. For example, given two sentences  $S$  and  $S'$ , it would predict whether  $S'$  proceeds  $S$  or not. However, RoBERTa drops the NSP objective, as Liu et al. (2019) observe that it worsened performance on downstream tasks. The hypothesis for this behavior is that the model was not able to learn long-range dependencies as well.

The original BERT model is pre-trained on a text corpus of 16GB, which the RoBERTa model increases tenfold.<sup>13</sup> Additionally, RoBERTa uses a richer encoding, namely a hybrid between character- and word level representations, in comparison to BERT’s character-level encoding. This encoding also has a larger vocabulary size, increasing it from 30 thousand to 50 thousand. Finally, RoBERTa is trained for substantially longer.<sup>14</sup>

On top of this RoBERTa architecture, I add a linear regression head, which takes in the encoding output of the model and applies a linear transformation to obtain a predicted score. Thus, given some candidate  $i$  and example sentence  $x_i$ , the output of this model can be written as

$$y_i = \beta \text{RoBERTa}(x_i)$$

where  $\text{RoBERTa}(x_i)$  outputs a  $d$ -dimensional vector and  $\beta$  represents the trainable parameters from the regression head. I fine-tuned this model to the task of predicting DW-Nominate scores, using the baseline of Congressional House members, as done in the MNIR approach. This fine-tuning process learns the optimal  $\beta$  coefficients as well as updating the existing weights in the transformer architecture by backpropagating the errors on the ideological prediction step through the entire model. For my implementation, I use all the default parameters for the HuggingFace

---

<sup>13</sup>In particular, BERT is trained on a text corpus of all Wikipedia articles and a large corpus of novels, BOOKCORPUS, for a total size of 16GB of text. The RoBERTa authors introduce 63 million new articles (CC-NEWS), web content extracted from URLs on Reddit (OPENWEBTEXT), and web content filtered to match the style of Winograd schemas (STORIES).

<sup>14</sup>While BERT is pre-trained for one million steps with mini-batches of 256 sequences, RoBERTa uses batch sizes of 8,000 sequences for 500,000 steps. This is roughly 15 times greater.

model `roberta-base`.<sup>15</sup> This model has 12 layers and 12 attention heads with a hidden dimension of 768, resulting in 125 million parameters. The model is trained over four epochs using a mean-squared error loss. Each epoch calculates out-of-sample validation loss.

I then use this model to calculate the fitted scores of the congressional candidates. Because the maximum token count for the RoBERTa model is 512, where a token corresponds to a unique word or sub-word, I aggregate each candidate's daily Tweets to serve as input to the model. This enables a much more finely grained measure of ideology than the other approaches.

One problem imposed by this methodology is that there is no clear way to weight observations by their ideological content. For classification tasks, the resulting probability can be interpreted as measuring the confidence of the classification; however, in the regression task, no such interpretation exists. In future work, I wish to consider how to adapt the architecture to include a weight estimate. One particular idea would be to adapt the RoBERTa architecture outlined above by replacing the current classifier with one that, given a text  $x_i$ , outputs a tuple  $(\hat{y}_i, \hat{w}_i)$  where  $w_i > 0$ . The objective would then be specified as  $\min w_i^2 (y_i - \hat{y}_i)^2$  for a true label  $y_i$ .

## 5. Data

### A. Candidate Metadata

My primary data source for the primary and general election candidates comes from Ballotpedia. *Ballotpedia* is a 501(c)3 charitable non-profit organization that contains write-ups for each congressional district race and biographies for all the participating candidates in both primary and general elections. All content is written by a “team of professional researchers, writers, and elections analysts.”<sup>16</sup> From these articles, I collect most of my district- and candidate-level metadata, including election data, such as primary dates and results, as well as a candidate's Twitter handle. I collect this information for all 470 congressional races in the 2020 election cycle, including Democratic and Republican primaries as well as all third-party candidates in the general elections. I exclude third-party primaries. Overall, my candidate sample includes 2,615 candidates, 924 of which are major-party general election candidates.<sup>17</sup> To construct my baseline sample, I gather all

<sup>15</sup>For a complete overview of the model and the configuration files, see <https://huggingface.co/roberta-base>.

<sup>16</sup>For more information about *Ballotpedia*, see the about section of the following website <https://ballotpedia.org/Ballotpedia:About>.

<sup>17</sup>The number of major-party candidates that ran in the primary and lost is 1,284. There are 358 third-party candidates.

of the Twitter handles from the 538 Senators and Representatives during the 116th Congress from Ballotpedia. These data were then joined with UCLA’s *Voteview* database containing information on all Members of the 116th Congress, including their calculated DW Nominate score (Lewis et al., 2022).

### B. Candidate Text Data

In order to get a measure of a candidate’s rhetoric, I use the Twitter API to pull all the tweets associated with the obtained candidate’s Twitter handle. For each tweet, I obtain its timestamp, its text, as well as auxiliary features such as the geolocation, number of likes, etc. I collect Tweets for all candidates and congressional representatives in my two samples. To ensure minimal overlap between the two samples, the baseline sample pulls from official government Twitter accounts while the candidate sample uses personal or campaign accounts whenever possible. Thus, taking Representative Alexandria Ocasio-Cortez as an example, the baseline sample would pull the Tweets associated with the account “@RepAOC” while the main sample would use tweets from her account “@AOC”. I am unable to identify a separate account for about 3% of the sample, for whom I use the official account.

Moreover, I choose the time periods for the two samples differently: for candidates, I collect tweets beginning eight months before the month of their primary and ending at the month of the general election while for the baseline sample, I use all tweets from the beginning of 2018 to February of 2020.<sup>1819</sup> Thus, there is minimal overlap between Tweets considered in the baseline sample and in the candidates samples, as desired. However, as a result of these decisions, I rely on the assumption that official Twitter accounts contain similar ideological content to personal accounts. Additionally, this approach must discard certain time-sensitive information; that is, issues that first arise during the general election may not be captured because they did not exist when the baseline sample was constructed. For example, the COVID-19 pandemic emerges during the election cycle but will not be in the baseline sample data.

---

<sup>18</sup>This baseline sample cutoff was intentionally chosen to avoid the height of the COVID-19 pandemic in the United States. The candidate sample predictions still are indirectly affected – for example, “public.health” is a major bigram in March of 2020 – but are otherwise better insulated from this event. Nonetheless, expanding the baseline sample to include the months of March and April 2020 does not change the findings, significantly. See the appendix for these results.

<sup>19</sup>The 116th Congress began on January 3, 2019 and ended on January 3, 2021. Thus, the first year of tweets (2018) hail from the members of the 116th Congress that were also in office during the 115th Congress. This is roughly 82% of the candidates in my baseline sample. I made the decision to expand the timeline in order to increase the number of bigrams in the baseline sample.

The vast majority of candidates in both of my samples maintain Twitter accounts. In particular, 81% of all candidates and 92% of all Democratic or Republican general election candidates have a Twitter account. For my baseline sample, I am able to collect Twitter handles for 99% of the Congress members. Over the entire course of the baseline period, I observe approximately 700,000 tweets with the average member tweeting about 1,350 times. For the candidate sample of two-party general election candidates I obtain around 800,000 tweets for an average of 778 tweets per candidate. Unfortunately, I am unable to ascertain the extent to which – if at all – candidates scrub their accounts, for example, by deleting certain tweets between the sent date and the date I query their handle.

### C. Descriptive Statistics

With the ideological predictions obtained from the different methodologies, I construct my final candidate dataset. Each row represents a unique candidate-month observation, which matches the time granularity for the MNIR and MFD methods. For RoBERTa I take the weighted average of each candidate’s predictions within a given month. Each row also contains candidate- and race-specific information, such as incumbency status, district partisan lean, Trump’s performance in the district, and election results.

For each candidate, my “panel” starts eight months before the primary election in their district and continues until the month of the general election. Thus, I standardize the period before the primary election, defining  $t = 0$  to the month of the primary for each candidate. Since the general election takes place in November for all races while state primaries occur in differing months, the number of periods after the primary varies by state. For a breakdown of which states share the same timelines, see Table 1. As an example, consider the 2020 Senate race in Oregon. The primary and general date were May 19 and November 3, respectively, a difference of approximately six months. Thus, candidates in this race would be observed for the eight months before the primary (labeled periods  $t \in \{-8, \dots, -1\}$ ), the month of the primary (May,  $t = 0$ ), and finally the five months until the month before the general ( $t \in \{1, \dots, 5\}$ ). I drop the month of November as it only contains two days before the general election. Additionally, candidates are dropped if they are missing from more than 25 percent of the time periods within their state’s panel. This threshold was chosen to ensure candidate’s are observed through the majority of the election cycle while also retaining the



size of the sample. However, my results are robust to this threshold (see appendix Tables A4-A6). My final dataset is summarized by Tables 2 and 3.

TABLE 1—STATES BY GENERAL ELECTION LENGTH

General Length (Months)	States
2 (Primary in Sept.)	DE, MA, MI, NH, RI
3	AK, AZ, CT, FL, HI, KS, MI, MN, MO, TN, VT, WA, WI, WY
4	ME, NJ
5	CO, GA, IA, ID, IN, KY, MD, MT, ND, NM, NV, NY, OK, PA, SC, SD, UT, VA, WV
6	NC, NE, OR
7	OH
8 (Primary in March)	AL, AR, CA, IL, MS, NC, TX

*Notes:* This table illustrates the breakdown of states according to the length of their general election. As the general election is always held in November, a general length of two implies that the primary was held in September and a length of eight would be in March. The largest clusters are for three months, five months, and eight months, making up 28%, 38%, and 14% of the states, respectively, and 18.4%, 27.5%, and 25.2% of the two-party candidates in my sample. I refer to these as short, medium, and long general elections, respectively.

I present descriptive statistics of selected candidate- and race-specific covariates, which will be used later on in my analysis, in Table 2. Incumbency status and party affiliation are presented as binary indicator variables, where a value of one corresponds to a candidate that is an incumbent or a Republican, respectively. In addition, district partisanship and competitiveness are detailed through the variables Competitive (Cook PVI), General Election Margin, and Trump 2020 Vote Share. The competitive indicator is obtained by selecting all districts with a Cook PVI rating within five points of neutral. The general election margin is reported as the positive difference between the winner’s share of the vote and the second placed candidate’s share, all on a scale from zero to one. Trump’s 2020 presidential performance in the district is expressed as a share. The number of months between the primary election and the general election is also reported.

Table 3 presents descriptive statistics at the candidate-monthly level for the ideology predictions across the different method as well as the underlying MNIR bigram and the MFD keyword counts. Statistics are reported by party. The MNIR and RoBERTa predictions represent the fitted DW-Nominate scores obtained from those methodologies while the MFD prediction corresponds to the scaled relative frequency of communal rhetoric. The counts section details the number of times a candidate employed a selected bigram in MNIR or a MFD Foundation keyword in a given month. Although I do not have data on the age of the candidates, there is no significant relationship

TABLE 2—SUMMARY STATISTICS OF CANDIDATE- AND DISTRICT-SPECIFIC COVARIATES

	Mean	SD	Median	N
<b>Candidate-Specific</b>				
Incumbent	0.474	0.5	0	665
Republican	0.439	0.497	0	665
Primary Election Competitive	0.068	0.252	0	665
Primary Election Margin	59.8	36.9	64.0	665
<b>District-Specific</b>				
Competitive (Cook PVI)	0.173	0.378	0	423
General Election Competitive	0.097	0.296	0	423
General Election Margin	28.1	20.8	23.7	423
General Election Length (Months)	5.26	2.06	5	423
Trump 2020 Vote Share	0.468	0.152	0.483	423

*Notes:* This table reports summary statistics for candidates and districts included in the final sample. Incumbent and Republican, are both binary indicators at the candidate-level for a candidate's incumbency status and party affiliation, respectively. Primary Election Margin details the margin between the first and second-place candidate in percentage points; Primary Election Competitive is a binary variable indicating if the margin fell within 7.5 percentage points. At the district level, Competitive (Cook PVI) is a binary indicator identifying if the district's Cook PVI rating was within a four point radius from even. General Election Margin reports the margin between the winning candidate and their opponent in percentage points; General Election Competitive is a binary variable indicating if this margin was within 5 percentage points. General Election Length reports the number of months in between the primary and the general election. Trump 2020 Vote Share is Trump's two-party vote share in the district. from the 2020 presidential election.

between age and word counts in the baseline sample.

These counts clearly indicate that Democratic candidates make greater use of Twitter. The median Democrat tweets about 3,000 more characters than the median Republican in a given month. This difference in tweeting habits explains the significantly higher counts observed among Democratic candidates than Republicans for both the MNIR and MFD methodologies. Indeed, these differences are roughly proportional: for example, the median Democrat tweets about 1.7 times as many characters and has 1.56 times as many bigrams.

TABLE 3—SUMMARY STATISTICS OF PREDICTIONS AND TEXT COUNTS, BY PARTY

	Democrat			Republican		
	Mean	Std. Dev.	Median	Mean	Std. Dev.	Median
<b>Predictions</b>						
MNIR	-0.188	0.361	-0.347	0.396	0.306	0.539
RoBERTa	-0.176	0.186	-0.212	0.216	0.223	0.251
MFD	-0.163	0.349	-0.168	0.051	0.366	0.022
<b>Counts</b>						
MNIR Bigrams	92	122	56	59.6	80.9	36
MFD Care	34.8	46.8	20	18.4	27.5	9
MFD Fairness	8.02	13.4	4	3.08	6.67	1
MFD Ingroup	27.4	36.4	16	17.4	26.3	9
MFD Authority	18.7	26.8	11	16	28	7
Tweet Length (Characters)	13,367	20,026	7,483	8,878	13,950	4,411

*Notes:* This table reports summary statistics for predictions obtained across the three methodologies as well as the bigram keyword counts in the final sample, split by party. Each observation is at the candidate-month level. For the predictions, MNIR and RoBERTa represent the predicted DW-Nominate scores from their respective models and MFD corresponds to the relative frequency of communal rhetoric. Counts are reported for the number of matching bigrams in MNIR as well as for the different MFD foundations. The length of the concatenated tweets in characters is also included.

Figure 3 plots the average MNIR bigram and MFD foundation counts per candidate over the course of the election cycle separately by party. The plot is normalized such that period zero represents the primary election for all races. As a result, candidates gradually drop out of the included averages, which is an unavoidable limitation of plotting the data over time given the varying general election lengths. Nevertheless, this figure still conveys that both MNIR and MFD counts are steadily rising through the primary and are substantially higher at any point in the general. Moreover, the relative percent change is roughly identical across until party.

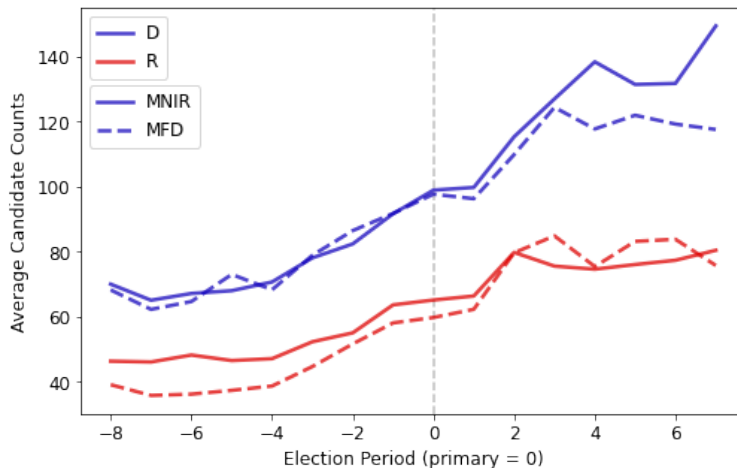


FIGURE 3. EVOLUTION OF MNIR AND MFD COUNTS, BY PARTY

*Notes:* This figure presents the average MNIR bigram (solid line) and MFD keyword (dashed) counts per candidate over the course of the election by party. This plot standardizes the primary period for all candidates. Because of the asymmetry in state primary timelines, candidates gradually drop out of the included average. For example, a candidate with only three months between the general and the primary would fall out of the sample at  $t = 2$ . Yet, it is clear that at all points in the general the counts for both parties are significantly greater than at the beginning of the primary.

Given I wish to measure ideological evolution and counts are increasing over the election cycle, one reasonable concern would be that my predicted extremity scores are a function of the number of counts that are observed, thereby confounding my estimates. For example, perhaps for lower counts, the model defaults to a moderate prediction as it has insufficient information. However, there is only a weak association between counts and within-in party ideology. The correlation coefficient between bigram counts and ideology is 0.09 and -0.23 for Republicans and Democrats, respectively, and for foundation counts it is even lower, at -0.01 and -0.07 for Republicans and Democrats. This suggests that a confounding relationship between counts and ideology should not be a problem. Nonetheless, all results using the MNIR and MFD estimates are weighted according to the number of counts. Additionally, the inclusion of the natural language model (RoBERTa), which is not dependent on word or phrase counts, should further bolster the robustness of these results.

While the variance of the predictions for each party is quite similar across the various methodologies, the center of the distribution varies substantially. For example, the MNIR method estimates that the median Democrat (-0.35) is 0.19 points more moderate in absolute terms than the median Republican (0.54). These differences are less apparent for the RoBERTa methodology as the range

of predictions is significantly compressed – perhaps due to greater regularization – as can be seen in Figure 4, and thus the median Democrat is estimated to be only 0.03 points more moderate. Given the bi-modal distribution of the DW-Nominate scores for Congress members, this suggests that the scale of the MNIR predictions is more accurate.

The correlation between the BERT and MNIR predictions over all the candidate-month observations is 0.63. Averaging the predictions over the entire electoral cycle, the correlation jumps up to 0.83, suggesting that the models may disagree with the exact timing of ideological shifts, but agree with the average ideological positioning of the candidate up to a scalar multiplier. These average predictions are displayed in Figure 4.

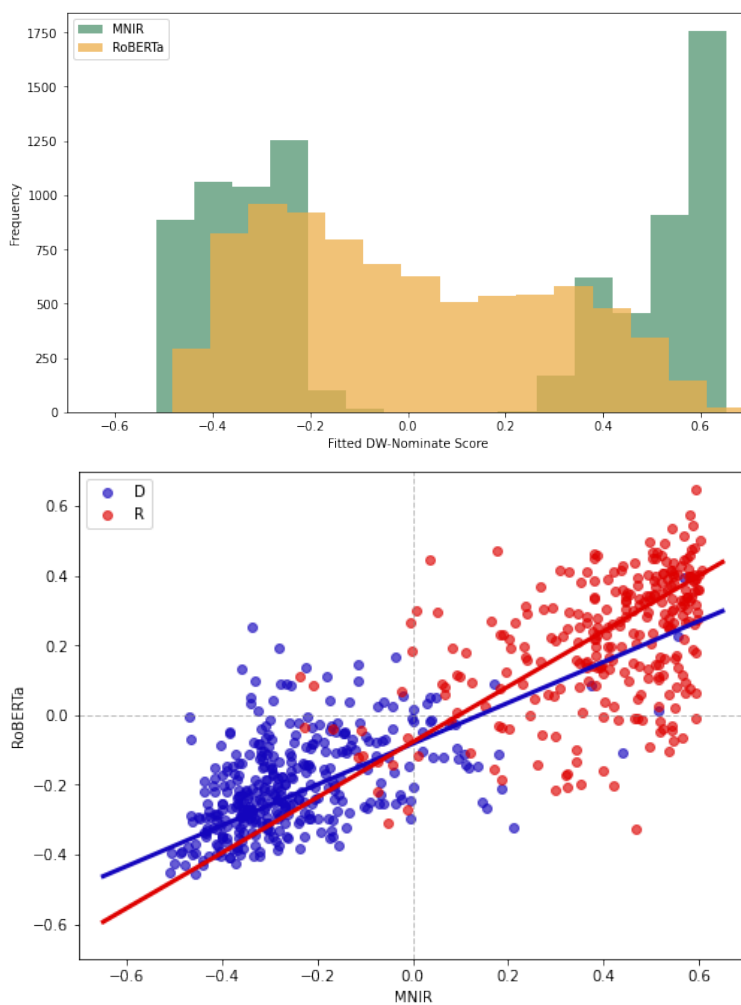


FIGURE 4. DISTRIBUTION OF DW-NOMINATE PREDICTIONS, MNIR AND ROBERTA

*Notes:* This figure presents the distribution of the obtained DW-Nominate 1 predictions from the MNIR and the RoBERTa methodologies for the candidate sample. The top plot depicts the distribution of candidate-month observation predictions for the two methods. The RoBERTa predictions are much more compressed and centered around zero, whereas the MNIR predictions better capture the bi-modal distribution of the true DW-Nominate scores (in Congress). The bottom plot shows the scatterplot of the average candidate predictions for the two methods, by party. In addition, the best-fit regression line is included. The correlation between the two measures is 0.83.

## 6. Validation

### A. Unveiling Model Behavior

In this subsection, I examine the textual features that the models are loading on. For the MNIR model, I first present the most partisan bigrams in Table 4 according to the  $\chi^2$  test, as described in the methods section, applied to both the baseline and the candidates sample. These results affirm intuition about the traditional policy aims of the two major parties, with Democratic candidates emphasizing gun violence, climate change, and health care and Republican candidates focusing on tax reform and growth, illegal immigration, and law enforcement. In addition, this table captures the extent to which the former President Trump and his administration captured the national political dialogue, with seven of the forty bigrams referencing him explicitly. However, Democrats and Republicans reference him differently: while Republicans refer to him according to his presidential title and tag him (“realdonaldtrump” was his Twitter handle), Democrats omit this honorific using either just his name or referencing the administration as a whole. Moreover, these partisan ideological themes appear to be fairly consistent across the two samples.

TABLE 4—MNIR MOST PARTISAN BIGRAMS, BY SAMPLE

Rank	Baseline		Candidate	
	Most Democratic	Most Republican	Most Democratic	Most Republican
1	gun_violence	tax_reform	health_care	president_realdonaldtrump
2	trump_administration	potus_realdonaldtrump	climate_change	nancy_pelosi
3	climate_change	president_realdonaldtrump	gun_violence	god_bless
4	health_care	speaker_pelosi	working_families	president_trump
5	pre_existing	adam_schiff	public_health	law_enforcement
6	background_checks	pro_growth	mitch_mcconnell	far_left
7	existing_conditions	great_news	voting_rights	radical_left
8	trump_admin	secure_border	affordable_care	thank_realdonaldtrump
9	voting_rights	born_alive	donald_trump	democrat_party
10	#forthepeople_pic	southern_border	social_security	men_women

*Notes:* This table presents the ten most Republican and Democratic bigrams in the baseline and candidate sample. This is calculated using the Chi-squared test outlined in the MNIR methodology section. Thus, these are the bigrams such that the null hypothesis that their usage is identical between Republicans and Democrats is most likely to be rejected. There is significant thematic overlap between the two samples, suggesting that the candidate sample generalizes well from the baseline.

Second, I calculate the predicted lift associated with each bigram, conducting the following thought experiment. Suppose that a candidate is observed tweeting a specific bigram, how would this affect the perception of their ideology – their other counts held constant? To address this, for

each bigram I modify each candidate-month bundle of counts to have zero and one occurrence of the given bigram and use the MNIR model and the forward regression to obtain ideological predictions for both states. I then define the lift as the average change in the predicted DW-Nominate scores among the candidates for each bigram. This allows me to identify which bigrams have the greatest impact in shifting ideological estimates, according to the MNIR model. I present these results in Table 5.

TABLE 5—BIGRAMS WITH THE GREATEST LIFT

Rank	Most Liberal	Most Conservative
1	free_school	protect_unborn
2	climate_science	democrats_think
3	profit_colleges	less_government
4	support_universal	life_pro
5	healthcare_right	radical_left
6	right_wing	pelosi_house
7	act_trump	trump_leadership
8	president_wants	born_alive
9	work_#forthepeople	corporate_interests
10	reproductive_rights	conservative_principles

*Notes:* This table presents the bigrams with the great lift according to the MNIR model. In order to compute this, for each bigram, I modify each candidate-month observation to have zero and one occurrence of this bigram and use MNIR to obtain a prediction for each case. I denote the average change over all observations as the lift. Thus, these bigrams represent bigrams whose inclusion would have the greatest affect on MNIR's predictions.

For the MFD approach, I calculate the most partisan keywords in the candidate sample according to a  $\chi^2$  test. The test is formulated identically to the test used in the MNIR approach, only with occurrences of keywords instead of bigrams. The results are presented in Table 6. As expected, keywords from the Authority foundation are most common among Republicans while the Fairness and Care keywords are more concentrated among Democrats. According to a  $\chi^2$  test on the overall foundations, Ingroup is the least partisan foundation and is quite similar across both parties.

Finally, for the RoBERTa approach, one of the main disadvantages of this deep neural architecture is that it is difficult to identify what features are most important. A common approach is to consider the self-attention scores as a metric for relevance, though this is further complicated by the fact that there are multiple attention heads and layers and no obvious procedure to aggregate these



TABLE 6—MFD MOST PARTISAN KEYWORDS IN THE CANDIDATE SAMPLE

Rank	Most Democratic		Most Republican	
	Keyword	Foundation	Keyword	Foundation
1	care	Care Virtue	riot+	Authority Vice
2	equal+	Fairness Virtue	communis+	Ingroup Virtue
3	law	Authority Virtue	patriot+	Ingroup Virtue
4	fight+	Care Vice	order+	Authority Virtue
5	communit+	Ingroup Virtue	illegal+	Authority Vice
6	justice	Fairness Virtue	destroy	Care Vice
7	rights	Fairness Virtue	terrorism+	Ingroup Vice
8	families	Ingroup Virtue	legal+	Authority Virtue
9	discriminat+	Fairness Vice	lawless+	Authority Vice
10	defen+	Care Virtue	caste+	Authority Vice

*Notes:* This table presents the ten most Republican and Democratic MFD keywords in the candidate sample. This is calculated using the Chi-squared-test as outlined in the MNIR methodology section, with the only difference being the use of MFD keywords instead of MNIR bigrams. The corresponding foundation for each keyword is also reported.

scores. Instead, I make use of a novel procedure to compute relevancy scores (Chefer, Gur and Wolf, 2021), which has shown improved performance across a range of NLP tasks. This method calculates relevancy scores for each attention head at each layer and backpropogates them through the network. These backpropogation gradients are used to aggregate the attention heads. I apply this method to two of the most partisan tweets according to the RoBERTa predictions and present the visualizations in Figure 5.

The first example presents an excerpt from Democratic Representative Don Beyer that received an estimated DW-Nominate score of -0.51 and the second from Republican Johsie Ezummadeen with an estimated score of 0.97. For the Republican tweet, the most relevant phrases appear to be “socialist” and “communist democrat,” though the model also picks up on the frequent mentions to Trump. Interestingly, a word like “vote,” which is likely commonly spoken across parties, is rightfully not selected by the model as important. The most important components of the Democratic tweet are significantly harder to determine. One major phrase appears to be “paid zero in ... taxes,” referencing a report by the New York Times that Trump paid no income taxes in 10 out of 15 years starting in 2000, an unmistakably partisan topic.<sup>20</sup> Otherwise, the uniform highlighting and lack of distinctly important phrases indicates that the model may perceive the

<sup>20</sup>The article is presented [here](#).

fundamental sentiment of the tweet to be left-leaning. This suggests that the model is succeeding at inferring partisanship from natural language, rather than just individual words or phrases.

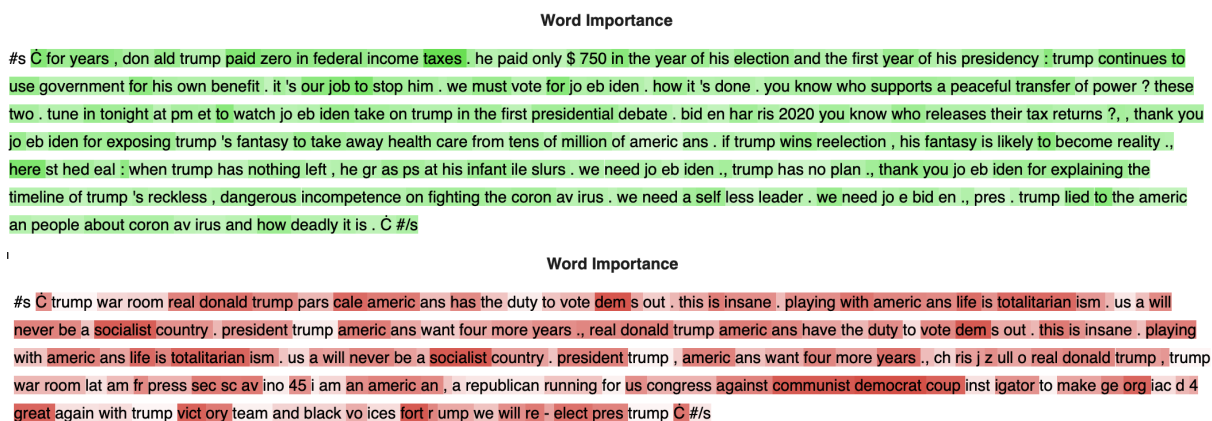


FIGURE 5. VISUALIZED RELEVANCY SCORES FOR ROBERTA

*Notes:* This figure depicts the relevancy scores obtained by running the procedure outlined in [Chefer, Gur and Wolf \(2021\)](#). The top row presents an example Tweet from Democratic Representative Don Beyer that had an estimated DW-Nominate score of -0.51. The bottom example uses a tweet from Republican Johsie Ezummadeen with an estimated score of 0.97. Both were in the top five most partisanly coded tweets in the candidate sample. Note, the results of this output are at the token level, though it is formatted for clarity.

### B. Out-of-Sample Validation

As discussed previously, the data for my baseline sample comes from the sitting House members in the 116th Congress. In this section, I assess how the models perform in-sample as well as out-of-sample, evaluating their fit on the subsample of Senators from this Congress. This is not a perfect sample for assessing external validity, particularly as it relates to the candidate sample, because the speech from the official Twitter accounts for Representatives is likely similar to that of Senators. However, since there do not exist gold-standard labels (such as DW-Nominate 1) for candidates without a roll-call record, this seems a natural choice.

I use the MNIR model and the specified forward regression to obtain the predicted DW-Nominate 1 scores of the sitting Representatives (in-sample) and Senators (out-of-sample) based on their word counts over the entire baseline period. I fit the forward regression according to three methods: first with linear regression, second with a boosted regression forest from `grf`, and third using gradient boosting from `xgboost`. The root-mean-squared error (RMSE) of the predictions from the three methods is 0.167, 0.090, 0.075 in-sample and 0.215, 0.143, and 0.131 out-of-sample. The

correlation with the true scores is 0.931, 0.980, 0.987 among the Representatives and 0.931, 0.949, and 0.957 among the Senators. In Figure 6, I visualize the distribution of the differences between the predicted and the true scores out-of-sample by the forward-regression method. The performance of the gradient boosting and regression forest methods is extremely similar while the linear regression performs substantially worse, particularly among more extreme Republicans (the last quartile in the plot). As a result, I selected the gradient boosting implementation for the rest of my analysis, though I include tables with the regression forest implementation in the appendix.<sup>21</sup>

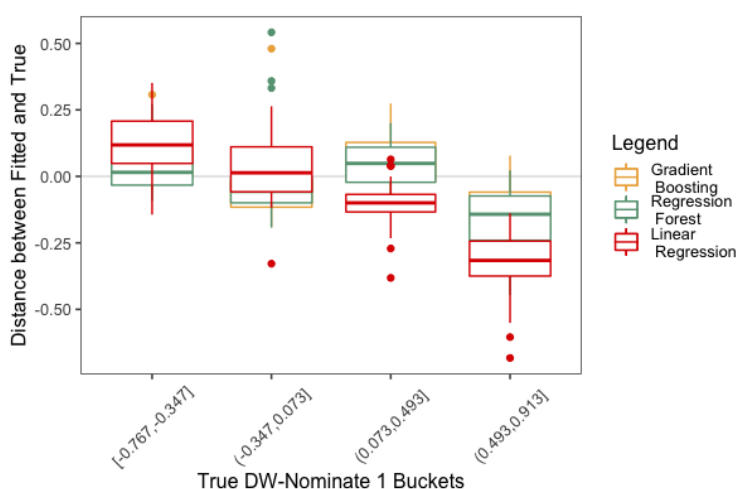


FIGURE 6. ERRORS ON SENATE FOR MNIR, BY FORWARD REGRESSION METHOD

*Notes:* This figure illustrates the distribution of errors on the Senate sample for the various MNIR prediction methods. In particular, I fit the forward regression with a gradient boosting implementation (yellow), a regression forest (green), and a linear regression (red). Linear regression has the largest errors, particularly at the extremes of the distribution. Gradient boosting and the regression forest perform quite similarly, likely due to the presence of a regularization term. Overall, the regression forest performed slightly worse than the gradient boosting approach.

Since RoBERTa has a maximum input sequence length of 512, it is not possible to evaluate the model over the entire period. Consequently, I aggregate the data at a daily level, concatenating each congressional members’s tweet for the given day. I then use the RoBERTa model to predict the ideological extremity of the candidate on each day and average these observations over the whole period. As discussed above, there is no clear way to weight the observations by their ideological content. Consequently, I weight the observations simply in accord with their character count, as a crude proxy for richness of content, penalizing especially short tweets. The averaged predictions yield a RMSE of 0.126 and 0.202 on the Representatives and the Seantors, respectively, and a

<sup>21</sup>For these results, see Table A3.

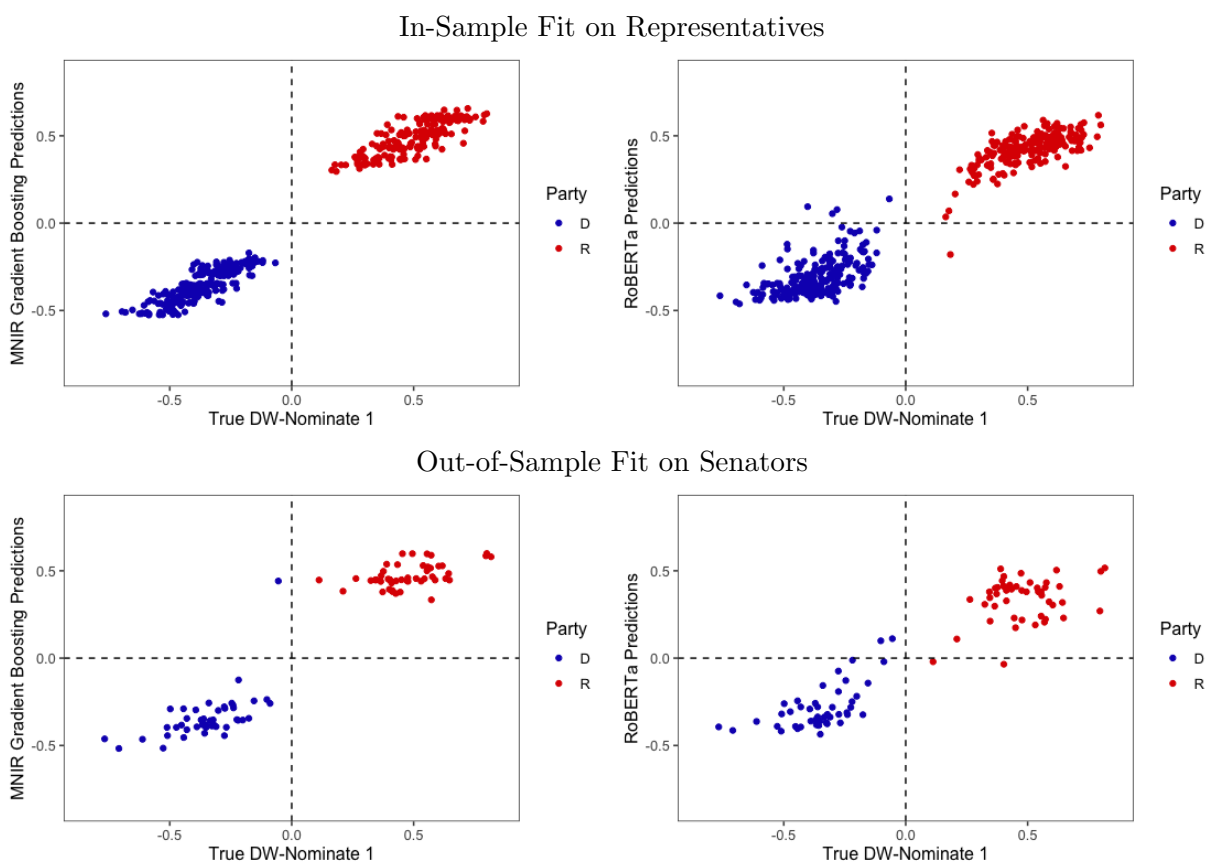


FIGURE 7. MNIR, ROBERTA FIT ON 116TH CONGRESS

*Notes:* This figure presents the results of the MNIR and RoBERTa predictions for the 116th Congress. Both models were trained on the House of Representatives members; these in-sample predictions are depicted in the top row. I then evaluated the out-of-sample fit of the model on the set of Senators. These results are shown in the bottom row.

correlation with the true scores of 0.971 and 0.926. Thus, the natural language approach performs slightly worse than MNIR. This is likely driven in part by the weighting issue. A comparison of these two methods is illustrated in Figure 7.<sup>22</sup>

Additionally, in order to assess the performance of the methods on the candidate sample, I compare the results among the winning candidates with their subsequent record as congressional representatives. In particular, for each winning candidate in my sample I calculate the average prediction over the entire election cycle and contrast it with their true DW-Nominate 1 score according to their roll-call votes in the 117th Congress. Although the fitted candidate scores are intended to capture shifts in ideological rhetoric – particularly, pandering away from the candidate’s true type – their overall election cycle average is an appropriate approximation of their true type. This analysis is limited in scope, as it only considers candidates that ultimately attain office, a sample that is of higher candidate quality.

TABLE 7—COMPARISON OF MODEL PREDICTIONS WITH 117TH CONGRESS SCORES

Model	MNIR		RoBERTa
	Gradient Boosting	Regression Forest	
Constant	−0.004 (0.009)	−0.015 (0.008)	0.010 (0.010)
Predictions	1.053*** (0.021)	1.161*** (0.022)	1.434*** (0.035)
<i>N</i>	326	326	326
<i>R</i> <sup>2</sup>	0.883	0.896	0.840

*Notes:* This table reports the results from regressing the true DW-Nominate score on the predicted DW-Nominate score from MNIR and RoBERTa among candidates in the 117th Congress. The true scores correspond to the candidate’s actual voting record after the election. The predicted scores are calculated using a weighted average over the course of the election cycle, where MNIR is weighted according to bigram counts and RoBERTa to the overall length of the text.

To analyze their accuracy, I specify a simple regression

$$\text{DW-Nominate 1 in 117th Congress}_c = \beta \text{Prediction}_c + \text{Intercept},$$

for a candidate  $c$ . For the Prediction variable, I use both the gradient boosting and regression

<sup>22</sup>A table with most partisan senators as estimated from MNIR and RoBERTa are included in appendix Tables A1 and A2.

forest MNIR predictions as well as the RoBERTa estimates. These results are displayed in 7. All three of these models are strongly correlated with the true scores: the MNIR predictions obtain an  $R^2$  of 0.883 and 0.896 across the two specifications with RoBERTa slightly lower at 0.84. The overall magnitude of estimates appears to be best calibrated in the gradient boosting MNIR implementation with values of  $\beta$  close to one and  $\alpha$  almost zero. With  $\beta = 1.43$ , the RoBERTa methodology clearly struggles with the magnitude of the predicted scores. Nonetheless, all of the models perform extremely well in capturing the relative differences between the candidates. Moreover, I train a forest-based model on the MNIR and RoBERTa scores together alongside a party indicator, and am able to further improve this  $R^2$  to 0.94. These findings provide support for the notion that the messaging during the election is not simply empty rhetoric and carries important ideological signal predictive of future voting records. In addition, they indicate that there is potential to use these predicted scores as a forecasting tool for candidate behavior, though this is left for future research.

## 7. Empirical Strategy

My empirical approach analyzes the extent of moderation in ideological rhetoric over the course of the election cycle and how race- and candidate-specific attributes interact with this movement.

First, I specify a difference-in-difference event study with indicator variables for each period. This has the additional benefit of clearly visualizing the evolution of candidate positioning over the course of the entire election cycle. Given the asynchrony in state primary election cycles, I split the sample according to the length of the general to ensure the timelines align. Thus, I estimate the following model:

$$(1) \quad y = X\theta^T + \sum_{t=0}^T \beta_{R,t} \text{Period } t \times \text{Republican} + \sum_{t=0}^T \beta_{D,t} \text{Period } t \times \text{Democrat}$$

for an election cycle with  $0, 1, \dots, T$  periods where  $T$  is the month before the general election and  $t = 0$  marks the initial period in my sample,  $y$  represents the outcome variable and  $X$  a matrix of covariates, such as incumbency status or district partisanship. Here,  $\beta_{R,t}$  and  $\beta_{D,t}$  represent the moderation in period  $t$  among Republicans and Democrats, respectively, relative to the initial period.

For the MNIR and the RoBERTa models, the dependent variable corresponds to the predicted DW-Nominate scores. For the MFD approach,  $y$  represents the scaled communal rhetoric scores. Since the range of these scores lies between -1 and 1 with Democrats clustered below and Republicans above the  $x$ -axis, the directionality of ideological moderation is flipped between the two parties. For example, when a Republican candidate  $i$  becomes more extreme,  $y_i$  increases toward one whereas for a Democratic candidate,  $y_i$  will decrease toward negative one. Thus, values of  $\beta_{R,t} < 0$  and  $\beta_{D,t} > 0$  capture moderation among the candidates.

This approach provides the finest time granulation and thus best reveals the timing of any ideological shifts. However, because the sample size is reduced by an approximate factor of five, this approach also has substantially higher variance and lower statistical power. Consequently, for my main specification I define a general election indicator variable, which allows me to return to the full sample of candidates. I estimate the following model:

$$(2) \quad y = \text{Intercept} + X\theta^T + \alpha \text{Republican} + \beta_{R,G} \text{General} \times \text{Republican} + \beta_{D,G} \text{General} \times \text{Democrat}.$$

Here, the intercept captures the extremity of Democrats in the primary,  $\alpha$  represents how much more or less extreme Republican candidates are than Democratic ones in the primary, and  $\beta_{R,G}$  and  $\beta_{D,G}$  represent the extent of moderation among Republicans and Democrats, respectively, in the general from the primary. Specifically, positive values of  $\beta_{R,G}$  and negative values of  $\beta_{D,G}$  indicate increasing extremity. Thus, in the case of the post-primary moderation hypothesis,  $\beta_{R,G}$  should be negative and  $\beta_{D,G}$  should be positive. To evaluate whether Republicans and Democrats moderate equally, I test the null hypothesis that  $\beta_{R,G} = -\beta_{D,G}$ .

In order to test for heterogeneity among the candidates' responses, I run the above specification with the addition of interaction terms based on candidate incumbency status and district competitiveness to form a triple difference-in-difference. Thus, for a binary interaction variable  $v$ , the

model becomes:

$$\begin{aligned}
 (3) \quad y &= \text{Intercept} + X\theta^T + (v \times X)\varphi^T + \alpha R + \gamma R \times v + \\
 &\quad \beta_{R,v,G} \text{ General} \times \text{Republican} \times v + \\
 &\quad \beta_{R,1-v,G} \text{ General} \times \text{Republican} \times (1 - v) + \\
 &\quad \beta_{D,v,G} \text{ General} \times \text{Democrat} \times v + \\
 &\quad \beta_{D,1-v,G} \text{ General} \times \text{Democrat} \times (1 - v).
 \end{aligned}$$

Here,  $\beta_{R,v,G}$  and  $\beta_{R,1-v,G}$  represent the moderation among Republican candidates with  $v = 1$  and  $v = 0$ , respectively, from the primary to the general, and  $\beta_{D,v,G}$  and  $\beta_{D,1-v,G}$  are defined analogously for Democratic candidates. Thus, in order to evaluate whether moderation is heterogeneous according to the interaction variable within party, I run the hypothesis test that  $\beta_{R,v,G} = \beta_{R,1-v,G}$  or  $\beta_{D,v,G} = \beta_{D,1-v,G}$ .

## 8. Results

### A. Main Results

Figure 8 presents the event study plots when using the MNIR and RoBERTa DW-Nominate fitted scores, respectively, separately grouped into short, medium, and long general elections. States with three months between the primary and general are classified as short, with five months as medium, and with eight months as long. These clusters contain 14, 19, and 7 states, accounting for 18.4%, 27.5%, and 25.2% of the two-party candidates in the sample, respectively. The long category comprises such a large share, as it includes California, Illinois, and Texas, three of the most populous states.<sup>23</sup> In the subsequent regressions, I introduce a binary general election indicator, enabling me to use the full sample of all states.

The plots are created by estimating the event study described in the empirical specification section and plotting the coefficients  $\beta_{R,t}$  and  $\beta_{D,t}$ . The MNIR regression is inverse-variance weighted by the number of bigram counts for each candidate-month observation while the RoBERTa regression is weighted according to the length of the input text. All regressions have standard errors clustered

<sup>23</sup>The states not included in these categories are distributed accordingly: five states with a difference of two months, two states with four months, three with six, and one with seven. See appendix Table 1 for a further breakdown of the states.



at the candidate level. 95% confidence intervals are included for the point estimates.

The six plots illustrate similar dynamics. Over the entire course of the election, there appears to be a trend of moderation among Republican candidates. In particular, the point estimates for the primary periods among Republicans are with few exceptions greater than the point estimates for the general periods in both figures. The majority of the plots appear to depict this shift beginning in early spring of 2020 and increasing over time until the end of the cycle. As the RoBERTa model has a finer time granularity, the shifts are more gradual whereas the MNIR model point estimates are more variable. Interestingly, the plots do not appear to identify a systematic change in the ideological rhetoric of the Democratic candidates. Many of the plots have relatively constant trend lines with slight increases or decreases in extremity. Finally, all of the plots estimate that Republican candidates are more extreme than Democratic candidates in magnitude throughout the election, though this appears to tighten in the general. However, these plots do not appear possible to identify a sharp discontinuity at the primary election where Republican candidates begin using more moderate rhetoric as suggested by the theoretical model. Rather the general shape of the plots seems to suggest a rapid increase in extreme rhetoric in the middle of the primary and then fairly gradual moderation thereafter. The consensus between the MNIR and RoBERTa model is encouraging, as it suggests that the different approaches are nonetheless identifying similar dynamics.

The event study plots have the benefit of depicting the potential nonlinear evolution of candidate positioning. However, since primary elections occur on different schedules, this visualization comes at the cost of statistical power, as the sample must be split up accordingly. Consequently, the confidence intervals on the individual period estimates are quite large, as can be seen in the plots. Indeed, while the trend of the point estimates clearly indicates moderation, the range of estimates included in the confidence interval signal that the null hypothesis of no moderation can also not be rejected, even among Republican candidates.

Since the event study results suggest that candidates moderate from the general to the primary but suffer from a lack of power due to the necessity of splitting the sample, my main results return to the full sample, as summarized in Tables 2 and 3. In order to use the full sample, I aggregate the periods using a binary indicator variable for the election period.

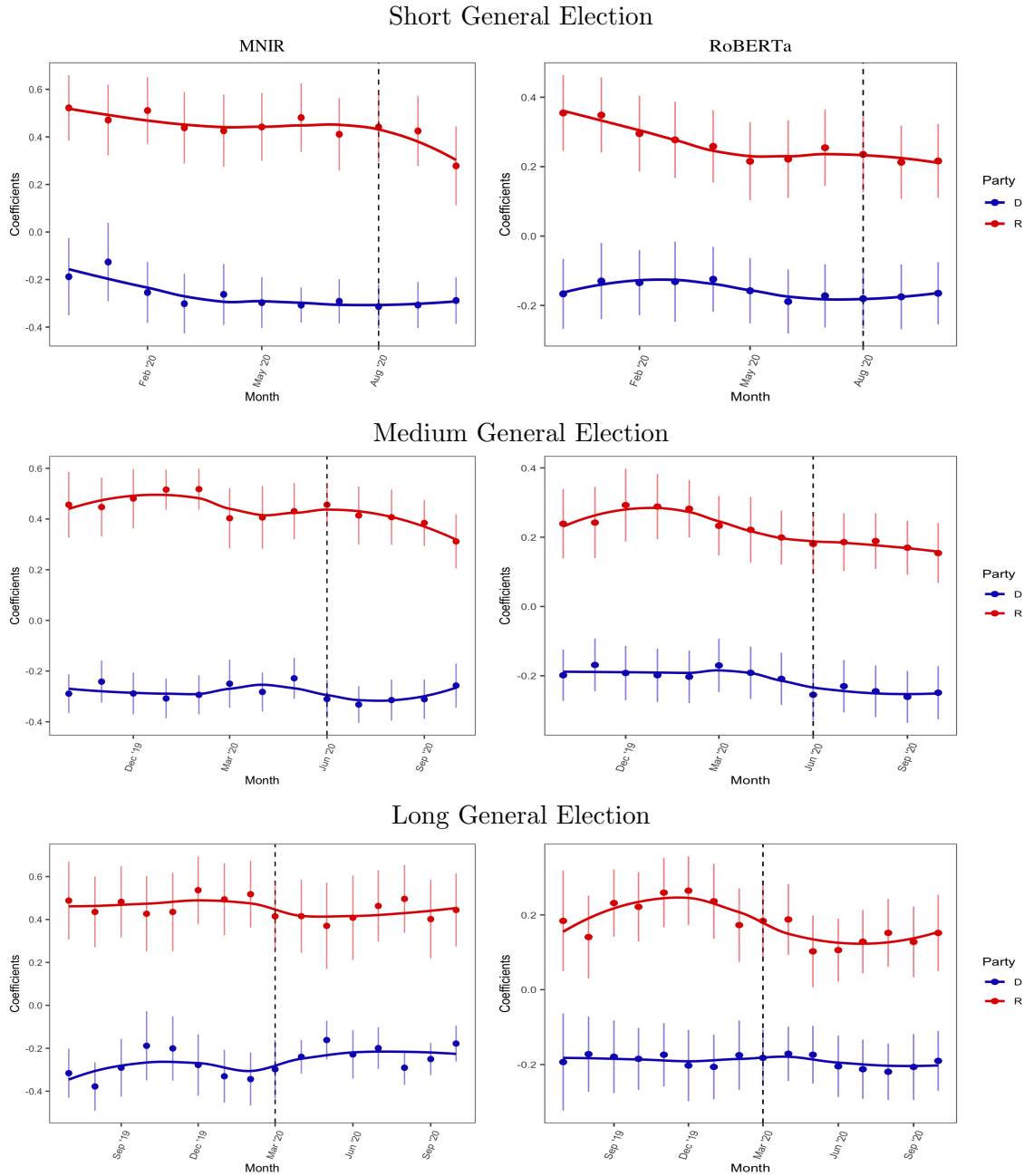


FIGURE 8. MNIR AND ROBERTA EVENT STUDY, BY GENERAL ELECTION LENGTH

*Notes:* This figure presents the event study plots obtained from estimating Equation 1 using the predictions obtained from the MNIR (left plots) and RoBERTa (right) methodologies. Since primary dates are determined at the state level, the length of the general election is variable. Consequently, the event studies are estimated on smaller subsamples of the full data with a shared election cycle. In particular, I use the following clusters: three month general or “short” (top plots), five month general or “medium” (middle), and eight month general or “long” (bottom). All specifications include a control for Trump’s 2020 presidential vote share in the district. Observations are weighted by the number of bigram counts for the MNIR results and the length of the tweet for RoBERTa. All standard errors are clustered at the candidate level.

I summarize the results of this main specification in Table 8 using the combined party sample with a general indicator, across the three ideology prediction methodologies. The first two columns of the table use predictions obtained by the MNIR model, the next two columns use the predictions from the pre-trained natural language model (RoBERTa), and finally the last two using the Moral Foundations Framework (MFD). All specifications use a binary general election indicator. The odd numbered columns also include President Trump’s share of the 2020 presidential vote as a covariate and the even numbered columns include additional controls for the candidate’s incumbency status, the chamber for which they are running, and the competitiveness of the race. All standard errors are clustered at the candidate level and observations are inverse-variance weighted.

The results in Table 8 show strong consensus across all three methodologies. All models estimate that there is not substantial change in rhetoric among Democratic congressional candidates, and that Republicans moderate significantly during the general election. These results affirm the findings from the event study. Moreover, the models are remarkably similar in the magnitude of their estimates. In Column 1 with Trump’s 2020 presidential vote share as a control, the MNIR model estimates that Republican candidates are 0.057 DW-Nominate 1 points more moderate in the general than the primary. With the inclusion of the additional controls in Column 2, the estimated effect is slightly larger, at 0.058 points. Similarly, RoBERTa identifies a 0.067 (Column 3) and 0.068 (Column 4) point decrease in extremity from the primary to the general. Finally, MFD estimates a 0.096 (Column 5) and 0.099 (Column 6) decrease in the usage of extreme univocalist or communalist rhetoric. These coefficients are significant at the 1% level for MNIR and at the 0.1% level for RoBERTa and MFD. The MNIR and RoBERTa estimates correspond to approximately half of a standard deviation in the distribution of DW-Nominate scores among House Republicans. This is roughly equivalent to half a standard deviation in the DW-Nominate distribution of congressional Republicans, akin to moving from the median Republican Senator Mike Crapo (0.51) to a slightly more moderate Senator like Cory Gardner (0.45) or Thom Tillis (0.43). On the extreme end, this movement is equivalent to moving from Joshua Hawley (0.73) to Rick Scott (0.66).

In stark contrast, all of the methodologies fail to find evidence of rhetorical movement among Democratic candidates over the course of the election cycle. Indeed, the estimated coefficients are small in magnitude and insignificant for almost all of the columns. Only the RoBERTa model with the full set of covariates identifies a statistically significant effect – at the 5% level – estimating

that Democratic candidates become more extreme by 0.029 points from the primary to the general. However, Democratic candidates are estimated to be more moderate overall in both the primary and the general than Republicans. For example, in a district where Trump had a 0.5 vote share, the average Democrat is estimated at to lie at -0.28 during the primary and the general. In contrast, the estimates for the Republican candidate in such a district are 0.47 in the primary and 0.42 for the general.

In addition, Table 8 includes the results of the linear hypothesis test for the null hypothesis that the extremity of rhetoric from Democrats and Republicans evolves identically from the primary to the general; that is, whether  $\beta_{R,G} + \beta_{D,G} = 0$ . For RoBERTa and MFD, the null hypothesis can be rejected at the 5% and 1% level, respectively. Although the MNIR estimates for Republicans are substantially larger than those for Democrats, the relatively large standard errors on the coefficients for Democrats result in the statistical insignificance of the MNIR difference.

These results demonstrate clearly that Democratic candidates did not engage in the same rhetorical movement as Republicans. This matches the party dynamics found in Burden (2001) with Democratic positions remaining constant over the two periods. One potential hypothesis for this discrepancy is that it has been suggested that the ideological centers of the Democratic primary base and general voters are closer than among Republicans (Grossmann and Hopkins, 2015). In this case, gains from extreme partisan signaling in the primary may be outweighed by the potential costs of appearing too extreme during the general election. Second, recent polarization trends have led to an asymmetry in the relative extremes of both parties, with the Republican party generally believed to be more extreme (McCarty, Poole and Rosenthal, 2016; Grossmann and Hopkins, 2015; Lewis et al., 2022). Thus, Democratic candidates may not need to stake out positions as extreme as Republicans to win over the base in the primary, and consequently face less of an incentive to moderate in the general. Alternatively, Democrats may be strategically reacting to the magnitude of their opponents' extremity, believing that they are still closer to independent voters even after their counterpart's pivot to the middle.

Finally, some have recently argued that the baseline DW-Nominate score is failing to accurately convey the different ideological cleavages within the Democratic party, thereby skewing the results of the prediction step. For example, DW-Nominate estimates Representative Alexandria Ocasio-Cortez – widely regarded as one of the members furthest on the Left – to be more conservative

TABLE 8—IDEOLOGICAL MODERATION AMONG 2020 CONGRESSIONAL CANDIDATES

	MNIR		RoBERTa		MFD	
	(1)	(2)	(3)	(4)	(5)	(6)
Republican × General	−0.057** (0.022)	−0.058** (0.021)	−0.067*** (0.012)	−0.068*** (0.012)	−0.096*** (0.019)	−0.099*** (0.019)
Democrat × General	0.015 (0.038)	0.014 (0.036)	−0.029* (0.013)	−0.029* (0.013)	−0.024 (0.014)	−0.029* (0.014)
Republican	0.740*** (0.019)	0.745*** (0.019)	0.410*** (0.019)	0.423*** (0.018)	0.269*** (0.019)	0.268*** (0.019)
Trump 2020 Vote Share	0.339*** (0.082)	0.336*** (0.076)	0.583*** (0.065)	0.618*** (0.070)	0.155* (0.062)	0.147* (0.062)
Incumbent		−0.012 (0.021)		0.039* (0.017)		−0.059*** (0.017)
Senate		−0.025 (0.024)		−0.019 (0.026)		−0.112*** (0.025)
Competitive		−0.044* (0.021)		−0.070*** (0.017)		−0.029 (0.018)
Constant	−0.455*** (0.037)	−0.433*** (0.032)	−0.473*** (0.033)	−0.487*** (0.041)	−0.156*** (0.033)	−0.102** (0.036)
Observations	8,349		8,304		8,349	
Candidates	665		661		665	
Outcome Mean	−0.056		−0.074		0.007	
Outcome SD	0.443		0.274		0.421	
$R^2$	0.609	0.612	0.576	0.593	0.096	0.107
Hypothesis Tests						
$\beta_{R,G} + \beta_{D,G} = 0$	−0.042 (0.045)	−0.044 (0.042)	−0.096*** (0.018)	−0.097*** (0.017)	−0.12*** (0.024)	−0.128*** (0.023)

*Notes:* This table presents the results from estimating Equation 2 on the final candidate sample for the different methodologies. The dependent variable is the predicted DW-Nominate scores from the MNIR model for Columns (1) and (2); the predicted DW-Nominate scores from the RoBERTa model for Columns (3) and (4); and the scaled relative frequency of communal rhetoric for Columns (5) and (6). Odd-numbered columns control only for Trump’s 2020 presidential vote share in the district; even-numbered columns also include indicators for incumbency status, competitiveness of the district (Cook PVI), and the congressional chamber. Observations are weighted by the number of bigram counts for the MNIR results, the length of the tweet for RoBERTa, and the number of keyword hits for MFD. All standard errors are clustered at the candidate level. In addition, the results of the hypothesis test on the equality of coefficients for the Republican and Democrat interaction terms are reported.

\* Significant at the 5% level.

\*\* Significant at the 1% level.

\*\*\* Significant at the 0.1% level.

than 90% of the Democratic Party. The same is true for many other members associated with the Democratic Party's progressive wing. (Lewis, 2022). This is an artifact of how DW-Nominate calculates ideological scores, as it does not consider the rationale behind the observed roll-call behavior. As a result, it is unable to distinguish between a progressive member voting no on a Democrat-sponsored bill for being too centrist from a Republican voting no for it being too liberal. If progressive Democrats are designated as more conservative in the baseline sample, both the MNIR and RoBERTa models would be inhibited in successfully identifying progressive rhetoric. This could in turn mask moderation among Democratic candidates. Nonetheless, specifying an adjusted DW-Nominate from Duck-Mayr and Montgomery (2021) measure that attempts to redress this problem by taking the directionality of a vote into account as the outcome variable did not alter this finding.<sup>24</sup> These results are included in the appendix for brevity, see Table A7. As a result, this explanation does not seem likely.

There are many possible explanations for the observation that candidates' moderation began prior to the conclusion of the primary election. First, most primaries are uncompetitive and thus the day the primary is unofficially decided may occur many months prior to the election date. Consequently, primary candidates may begin preparations for the general election, such as moderating rhetoric, well in advance of the primary election. For example, a candidate's internal polling many months before the election might show them comfortably leading the primary race but depict a relatively close general election, in which case the candidate might begin strategically positioning themselves to attract more general election voters immediately. Or, alternatively, a candidate's primary competitor might drop out of the race, cementing the victory instantly.

Second, candidates may fear changing rhetoric too drastically or suddenly, thereby risking accusations of insincerity and inauthenticity. These accusations, described above as flip-flopping costs, are commonplace in American political media and likely are incorporated in a candidate's calculus. Thus, candidates may choose to slowly alter their rhetoric, beginning this change during the primary before general election voters are paying attention. Finally, this could reflect exogenous changes in the national environment. In particular, major political events have the potential to shift the national discussion, forcing candidates to respond to these issues, while rhetoric on these topics

---

<sup>24</sup>This model is a Generalized Graded Unfolding model which is estimated via a Metropolis coupled Markov chain Monte Carlo approach, and is referred to as MC3-GGUM. For more information, see Duck-Mayr and Montgomery (2021).

might be partisanly coded, thereby altering the perceived ideological content of their rhetoric.

As mentioned above, the MNIR plots appear to reflect a large moderating effect around March and April of 2020, which is then sustained for the rest of the cycle. This time period captures the height of the COVID-19 pandemic during the election cycle. Since the baseline sample does not overlap with this period, ending in February of 2020, none of the bigrams explicitly reference COVID. This is both an important benefit and limitation of this methodology, as it better isolates it from exogenous events, though how candidates discuss important events may also be indicative of ideology. Yet, it is still likely that this period elevated certain topics to the national forefront, which are contained in the selected set of bigrams. Indeed, this can be seen by analyzing the most common bigrams during this time period, with bigrams like “small\_businesses” – the most common bigram among Republicans in March – and “public\_health” or “health\_care” – the ninth and tenth most common among Republicans – vaulting to the forefront. Nonetheless, the popularity of health-related terms quickly declined while the observed moderation is maintained. For example, “public\_health” falls from the ninth to the 17th most common bigram among Republicans from March to April, and by May, it is only the 64th. Moreover, the direction of these effects is not necessarily identical – while the coefficients of the MNIR model suggest that “public\_health” is a more left-leaning phrase, “small\_businesses” is identified as having right-leaning implications. Thus, the emergence of COVID-19 is insufficient to explain the sustained moderation observed among Republican candidates, though it may have acted as a catalyst to spur this movement.

The consensus across these three categorically different methodological approaches provides solid evidence that the ideological rhetoric of Republican candidates evolved over the course of the election cycle, and in particular was more moderate in the general election than the primary election. The effect identified is also large, at approximately half of a standard deviation in ideology for the Republican House Caucus. While these results do lend support for the post-primary moderation hypothesis, there are important reasons to be cautious in analyzing the underlying reasons behind these rhetorical changes. As this paper only collected data for congressional races in the 2020 election year, the analysis is unable to control for the possible underlying political environment. Consequently, it is possible that the national environment was globally shifting in the direction of the Democrats, and thus Republicans were forced to moderate in reaction. To address this concern, I proceed by analyzing races within this election year for which there should be additional incentives

for strategic moderation.

### B. Heterogeneity Results

This section considers the potential heterogeneity in the evolution of ideological rhetoric based on candidate- and race-specific characteristics. In particular, I analyze how incumbency status and both primary and general election competitiveness interact with moderation. These results are reported in Tables 9, 10, and 11.

#### INCUMBENCY STATUS

Applying the logic of asymmetric prominence as outlined in the model section, because of the previous political history that an incumbent has, it should be more difficult for the candidate to signal to voters to distinguish their true type. Consequently, incumbency status should diminish a candidate's ability to moderate. In addition, voters may also impose penalties on candidates who noticeably engage in "flip-flopping" behavior, which should further reduce moderation (Hummel, 2010). Voters are thought to punish "flip-flopping" as it impacts their judgments of the candidate's trustworthiness and reliability (Tomz and Van Houweling, 2009). Nevertheless, incumbents are also more experienced and likely of higher quality than non-incumbents, which may enable them counteract some of these costs.

The results using incumbency as an interaction are reported in Table 9. Each regression uses Trump's vote share as a control, as in the odd-numbered columns in Table 8. In addition, hypothesis tests on whether incumbent candidate's moderate as much as non-incumbent candidates are included as well (for Republicans:  $\beta_{R,I,G} - \beta_{R,NI,G} = 0$ ; and for Democrats:  $\beta_{D,I,G} - \beta_{D,NI,G} = 0$ ).

The results in Table 9 do not suggest that there are substantial differences in moderation by incumbency status. In particular, MNIR estimates that incumbents and non-incumbents moderate by 0.058 points. Neither the estimate on non-incumbents nor the difference between the two is statistically significant. RoBERTa estimates incumbents and non-incumbents to moderate by 0.05 and 0.077 points, respectively, a slightly larger difference of 0.027 points that is also not significant. However, the MFD approach suggests a substantial difference between the two estimates, with incumbents moderating by 0.134 and non-incumbents by 0.066. This is a difference of 0.068, which is significant at the 10% level. In addition, MFD suggests that Democratic non-incumbents actually



get more extreme by 0.047 points in the general, though this is not a significant difference.

These results do not provide clear evidence to support the notion that incumbents moderate less than non-incumbent candidates. One potential reason for why the additional costs imposed on incumbents to moderate are not binding is that congressional races are not high profile enough. As discussed above, given that incumbents have extensive political histories, they were hypothesized to face a greater asymmetric prominence, inhibiting their ability to moderate. However, if the race itself is not prominent enough – perhaps as Americans are increasingly consuming national news (Hayes and Lawless, 2018) or because the race itself is not competitive – voters may not be cognizant of these shifts. Alternatively, it is possible that consequences induced from flip-flopping are simply offset by various other factors, such as candidate experience and quality – that is, that incumbents are potentially more practiced and tactful in strategically altering their rhetoric. Finally, it is possible that the estimate for non-incumbents is pushed downward, as it contains more candidates in uncompetitive elections with a minimal chance of victory and thus little reason to moderate. Consequently, in the next section, I analyze the effects of general election competitiveness.

#### GENERAL ELECTION COMPETITIVENESS

Applying the logic of strategic moderation in the post-primary moderation hypothesis, the general election competitiveness of a race should accentuate the movement observed. In the language of Agranov’s model, these races have an especially high prominence, and consequently, moderate candidates have a greater chance of victory in the general and will thus expend more effort to pander in the primary. These dynamics suggest that candidates in competitive generals will likely moderate more than those in uncompetitive elections. Additionally, independent voters are especially crucial to winning the election in competitive elections, thus further incentivizing the candidates to cater toward this more moderate set of voters. In order to address this hypothesis empirically, I consider two measures of general election competitiveness. First, I construct a binary indicator variable based on the Cook PVI rating, which quantifies the lean of each congressional district based on the historical presidential performance in the district relative to the nation as a whole. In particular, I define all districts that are within five points of the national average (between +5D and +5R) according to Cook PVI as competitive districts. Second, I use the actual general election results, defining all the races that had an eventual general margin within five percentage points as

TABLE 9—IDEOLOGICAL MODERATION, BY INCUMBENCY STATUS

	MNIR	RoBERTa	MFD
	(1)	(2)	(3)
Republican × Incumbent × General	−0.058** (0.018)	−0.050*** (0.013)	−0.134*** (0.032)
Republican × Non-Incumbent × General	−0.058 (0.032)	−0.077*** (0.016)	−0.066** (0.021)
Democrat × Incumbent × General	−0.009 (0.013)	−0.022** (0.008)	−0.008 (0.021)
Democrat × Non-Incumbent × General	0.031 (0.064)	−0.031 (0.020)	−0.047** (0.018)
Republican	0.736*** (0.026)	0.374*** (0.034)	0.265*** (0.025)
Incumbent	−0.021 (0.062)	0.068 (0.080)	−0.059 (0.076)
Republican × Incumbent	0.009 (0.035)	0.168*** (0.046)	−0.009 (0.051)
Trump 2020 Vote Share	0.293** (0.098)	0.552*** (0.124)	0.163 (0.088)
Incumbent × Trump 2020 Vote Share	0.034 (0.125)	−0.236 (0.145)	0.017 (0.164)
Constant	−0.430*** (0.050)	−0.454*** (0.074)	−0.130* (0.054)
Observations	8,349	8,304	8,349
Candidates	665	661	665
Incumbents	315	313	315
$R^2$	0.609	0.591	0.102
Hypothesis Tests			
$\beta_{R,I,G} - \beta_{R,NI,G} = 0$	0.00 (0.037)	0.028 (0.020)	−0.068 (0.039)
$\beta_{D,I,G} - \beta_{D,NI,G} = 0$	−0.04 (0.065)	0.009 (0.022)	0.039 (0.028)

*Notes:* This table presents the results from estimating Equation 3 with an indicator for incumbency status on the final candidate sample for the different methodologies. The dependent variable is the predicted DW-Nominate scores from the MNIR model for Column (1); the predicted DW-Nominate scores from the RoBERTa model for Column (2); and the scaled relative frequency of communal rhetoric for Column (3). All specifications control for Trump's 2020 presidential vote share in the district. Observations are weighted by the number of bigram counts for the MNIR results, the length of the tweet for RoBERTa, and the number of keyword hits for MFD. All standard errors are clustered at the candidate level. The results of hypothesis tests on the equality of the coefficients for Republican and for Democratic incumbents vs. non-incumbents in the general election are reported.

\* Significant at the 5% level.

\*\* Significant at the 1% level.

\*\*\* Significant at the 0.1% level.

competitive. From hereon, I will refer to the first and second as the Cook PVI and general election margin definition of competitiveness, respectively.

These definitions have tradeoffs. For example, the Cook PVI definition is exclusively based on past performance and does not account for candidate-specific advantages. However, because of this, it is also a measure that best reflects the average tendency of the district, unbiased by the current congressional campaigns. In contrast, the general election margin is likely in part affected by the rhetorical choices of the candidates – for example, the political science literature suggests more moderate candidates perform better in general elections (Hall, 2015) – and thus, this definition of competitiveness may be biased. Nonetheless, the general election margin definition is much narrower, focusing only on races with an eventual margin of five points, whereas the average margin in competitive districts based on the Cook PVI definition is 11.65 percentage points. Consequently, the Cook PVI definition is more expansive, including 95 districts, compared to the 41 in the general election margin definition.

The first three columns in Table 10 use the Cook PVI definition and the last three use the general election margin definition. Each regression uses Trump’s vote share as a control, as in Table 9. In addition, I include hypothesis tests for the whether Republicans and Democrats moderate significantly differently in competitive versus non-competitive districts ( $\beta_{R,C,G} - \beta_{R,NC,G} = 0$  and  $\beta_{D,C,G} - \beta_{D,NC,G} = 0$ , respectively).

MNIR estimates a 0.134 and 0.169 point moderation among Republican candidates in competitive districts according to the Cook PVI and general election margin definitions, respectively. These are 0.109 and 0.138 points larger than the estimates for candidates in non-competitive districts. The first difference is significant at the 5% level, and the second at the 1% level. Both MFD and RoBERTa estimate a larger moderating effect among Republicans in competitive than non-competitive races according to both definitions, though these differences are substantially smaller than in MNIR. In particular, RoBERTa and MFD estimate that Republicans in competitive elections according to the Cook PVI definition moderate by -0.026 and -0.027 more than those in non-competitive generals. According to the general election margin definition, the differences are even smaller, at -0.006 for RoBERTa and -0.013 for MFD. None of these differences are significant, however. Thus, there is mixed support for the hypothesis that Republican candidates moderate more in competitive elections.

TABLE 10—IDEOLOGICAL MODERATION, BY GENERAL ELECTION COMPETITIVENESS

	Cook PVI			General Margin		
	MNIR (1)	RoBERTa (2)	MFD (3)	MNIR (4)	RoBERTa (5)	MFD (6)
Republican × Competitive × General	-0.134** (0.042)	-0.088*** (0.017)	-0.117*** (0.029)	-0.169*** (0.035)	-0.073** (0.023)	-0.108** (0.037)
Republican × Non-Competitive × General	-0.025 (0.021)	-0.063*** (0.015)	-0.089*** (0.023)	-0.031 (0.025)	-0.067*** (0.013)	-0.095*** (0.021)
Democrat × Competitive × General	-0.007 (0.016)	-0.056*** (0.015)	-0.081* (0.037)	0.002 (0.027)	-0.066*** (0.016)	-0.078 (0.043)
Democrat × Non-Competitive × General	0.021 (0.048)	-0.024 (0.015)	-0.013 (0.015)	0.019 (0.043)	-0.025 (0.014)	-0.017 (0.015)
Republican	0.761*** (0.022)	0.430*** (0.023)	0.254*** (0.022)	0.736*** (0.021)	0.408*** (0.022)	0.262*** (0.021)
Competitive	-0.225 (0.188)	-0.284 (0.166)	-0.126 (0.259)	-0.793** (0.298)	-0.313 (0.306)	-0.155 (0.397)
Republican × Competitive	-0.065 (0.043)	-0.045 (0.037)	0.064 (0.046)	0.029 (0.041)	0.013 (0.042)	0.055 (0.057)
Trump 2020 Vote Share	0.323*** (0.083)	0.575*** (0.064)	0.164** (0.062)	0.338*** (0.083)	0.583*** (0.065)	0.164** (0.062)
Competitive × Trump 2020 Vote Share	0.476 (0.409)	0.528 (0.352)	0.199 (0.552)	1.583** (0.613)	0.649 (0.620)	0.226 (0.836)
Constant	-0.447*** (0.038)	-0.463*** (0.033)	-0.153*** (0.034)	-0.453*** (0.037)	-0.473*** (0.033)	-0.155*** (0.033)
Observations	8,349	147,666	8,349	8,349	147,235	8,349
Candidates	665	661	665	665	661	665
Candidates in Competitive Races	136	135	136	76	76	76
$R^2$	0.615	0.59	0.098	0.612	0.579	0.097
Hypothesis Tests						
$\beta_{R,C,G} - \beta_{R,NC,G} = 0$	-0.110* (0.047)	-0.026 (0.023)	-0.027 (0.037)	-0.138** (0.043)	-0.006 (0.027)	-0.013 (0.042)
$\beta_{D,C,G} - \beta_{D,NC,G} = 0$	-0.027 (0.051)	-0.033 (0.021)	-0.068 (0.040)	-0.017 (0.050)	-0.042 (0.022)	-0.060 (0.046)

Notes: This table presents the results from estimating Equation 3 with an indicator for general election competitiveness on the final candidate sample for the different methodologies. For Columns (1)-(3), this indicator identifies candidates in districts with a Cook PVI rating within a four point radius from “EVEN”; and for Columns (4)-(6), candidates in districts with a final general election margin within five percentage points. The dependent variable is the predicted DW-Nominate scores from the MNIR model for Columns (1) and (4); the predicted DW-Nominate scores from the RoBERTa model for Columns (2) and (5); and the scaled relative frequency of communal rhetoric for Columns (3) and (6). All specifications control for Trump’s 2020 presidential vote share in the district. Observations are weighted by the number of bigram counts for the MNIR results, the length of the tweet for RoBERTa, and the number of keyword hits for MFD. All standard errors are clustered at the candidate level. The results of hypothesis tests on the equality of the coefficients for Republican and for Democratic candidates in the general election from competitive vs. non-competitive generals are reported.

\* Significant at the 5% level.  
 \*\* Significant at the 1% level.  
 \*\*\* Significant at the 0.1% level.

Among Democrats, the RoBERTa estimates suggest that candidates in competitive elections actually become more extreme in the general than their counterparts in uncompetitive races. In particular, they are estimated to become more extreme by 0.056 and 0.066 for the Cook PVI and general election margin definitions, respectively. This implies 0.032 and 0.041 points in increasing extremity relative to Democrats in non-competitive elections, though the estimates for this latter group are not significant. The second difference is just shy of significance at the 5% level. Similarly, MFD shows a relatively large increase in extremity among Democratic candidates in competitive districts. However, these coefficients are also subject to quite large standard errors, and as a result, these differences are not significant. MNIR does not identify an effect across either specification.

Clearly, there is substantial divergence among the different methodologies. Nonetheless, the MNIR results among Republican do provide suggestive evidence that candidates engage in strategic moderation, which is particularly accentuated in prominent, competitive environments. In contrast, the effect identified in the RoBERTa model of Democrats getting more extreme in these districts is a note for caution. Given that my data only extends to one year and the sample of candidates running in competitive races is small – only 76 total candidates are included in the general election split – these results are likely fairly sensitive to individual trends or noise in these underlying districts. Moreover, the limited sample substantially reduces the power of the regression and helps explain why, despite the point estimates being fairly divergent, the difference is often not significant. Accordingly, I conclude that results among Republican candidates lend cautious support to the hypothesis that the extent of moderation is increased in competitive general elections.

#### PRIMARY ELECTION COMPETITIVENESS

The primary election competitiveness of a race should further intensify the level of moderation over the course of the election cycles. One major component of the post-primary moderation hypothesis is that candidates are dragged to the extremes in the primary in order to appease the base. These centrifugal forces should be particularly strong in competitive primaries. This result was also demonstrated in Agranov's model, where the heightened prominence of a primary makes signalling more effortful for moderate type candidates. To attain the nomination in a competitive primary, candidates consequently likely have to take on more extreme positions, from which they

need to recover in order to persuade general election voters. In order to empirically test whether primary election competitiveness increases moderation, I define a measure of competitiveness according to the eventual primary election margin between the top two candidates. In particular, I define candidates to have run in a competitive primary election if they won by five percentage points or less. These results are reported in Table 11.

None of the methodologies identify a significant effect of additional moderation in this sample. MNIR does not obtain a significant coefficient for Republicans or Democrats in competitive primaries in the general. Additionally, RoBERTa and MFD find slight differences among Republicans in these two samples, though in conflicting directions. RoBERTa estimates that Republican candidates in competitive primary elections only moderate by 0.04 points, which is 0.03 points less than those in uncompetitive elections. In contrast, MFD finds that Republicans in the competitive primary sample moderate by 0.023 more points. However, the large standard errors on the coefficient for Republicans in competitive primaries ensures that these differences are not significant. Moreover, no significant effect is observed among Democrats across all methodologies.

One reason potentially explaining the lack of difference between these candidates is that races with competitive primaries may not necessarily have a competitive general. Thus, there may not exist a strong incentive to moderate because the median voter in that district is already fairly partisan. In fact, it seems reasonable that an extremely one-sided district may have competitive primaries, as candidates understand that upon winning the primary the general is all but guaranteed. This hypothesis is borne out by the data, as the mean general election margin for these races is 24.38% with a 17.05% standard deviation. In such a scenario, there would be no incentive to moderate in the general. Additionally, as discussed in the general election competitiveness section, the number of candidates included in these definitions is small, leading to significantly reduced power. Here, only 46 candidates faced a competitive primary. Consequently, a lot is being loaded on the individual candidates within this subset. In order to get more precise estimates, future work should aim to collect data from multiple election years.

TABLE 11—IDEOLOGICAL MODERATION, BY PRIMARY ELECTION COMPETITIVENESS

	MNIR	RoBERTa	MFD
	(1)	(2)	(3)
Republican × Competitive Pri × General	0.033 (0.061)	−0.040* (0.019)	−0.118* (0.059)
Republican × Non-Competitive Pri × General	−0.065** (0.022)	−0.070*** (0.013)	−0.095*** (0.020)
Democrat × Competitive Pri × General	0.014 (0.035)	−0.058 (0.035)	−0.040 (0.038)
Democrat × Non-Competitive Pri × General	0.014 (0.039)	−0.027 (0.014)	−0.023 (0.015)
Republican	0.741*** (0.020)	0.411*** (0.021)	0.264*** (0.020)
Competitive Primary	0.119 (0.104)	0.109 (0.062)	0.098 (0.098)
Republican × Competitive Primary	−0.017 (0.070)	−0.043 (0.049)	0.049 (0.074)
Trump 2020 Vote Share	0.365*** (0.084)	0.598*** (0.070)	0.170** (0.064)
Competitive Primary × Trump 2020 Vote Share	−0.318 (0.217)	−0.170 (0.120)	−0.168 (0.196)
Constant	−0.466*** (0.038)	−0.482*** (0.036)	−0.164*** (0.034)
Observations	8,323	147,117	8,323
Candidates	665	661	665
Candidates in Competitive Primaries	46	46	46
$R^2$	0.612	0.576	0.096
Hypothesis Tests			
$\beta_{R,CP,G} - \beta_{R,NCP,G} = 0$	0.098 (0.065)	0.029 (0.023)	−0.023 (0.062)
$\beta_{D,CP,G} - \beta_{D,NCP,G} = 0$	0.00 (0.052)	−0.031 (0.038)	−0.017 (0.041)

*Notes:* This table presents the results from estimating Equation 3 with an indicator for primary election competitiveness status on the final candidate sample for the different methodologies. This indicator identifies candidates with a primary election margin within 7.5 percentage points. The dependent variable is the absolute value of: the predicted DW-Nominate scores from the MNIR model for Column (1); the predicted DW-Nominate scores from the RoBERTa model for Column (2); and the scaled relative frequency of communal rhetoric for Column (3). All specifications control for Trump’s 2020 presidential vote share in the district. Observations are weighted by the number of bigram counts for the MNIR results, the length of the tweet for RoBERTa, and the number of keyword hits for MFD. All standard errors are clustered at the candidate level. The results of hypothesis tests on the equality of the coefficients for Republican and for Democratic candidates in the general election from competitive vs. non-competitive primaries are reported.

\* Significant at the 5% level.

\*\* Significant at the 1% level.

\*\*\* Significant at the 0.1% level.

## 9. Conclusion

Taking advantage of the proliferation of social media in political communication, I collect candidate speech on Twitter over the course of the 2020 congressional election cycle in order to investigate the extent of moderation from the primary to the general. Using three different approaches, I translate these text data into estimates of ideological extremity. First, I take a data-driven approach to selecting the most partisan bigrams and specify a Multinomial Inverse Regression to predict ideology. Second, I use a theoretically-derived set of keywords to construct a measure of the frequency with which candidates invoke moral values associated with political convictions. Finally, I specify a natural language model using a deep learning model, which I fine-tune on the task of ideological prediction. With these estimates, I am able to construct a novel dataset consisting of almost every 2020 congressional candidate with monthly estimates of their ideological positioning.

I then validate these measures, demonstrating that they perform well out-of-sample. In particular, I evaluate my models on the set of Senators from the 116th Congress as well as on the set of candidates in my sample that won their race in the 2020 election and thus participated in the 117th Congress. In both cases, my best model's predictions capture almost 90% of the variance in the true DW-Nominate 1 scores. Additionally, the models successfully identify ideologically meaningful language.

Using these measures, I document a systematic moderating trend among Republican candidates across all methodologies consistent with the post-primary moderation hypothesis. This movement is approximately equivalent to half of a standard deviation in the baseline sample of House Republicans. However, I do not find a significant shift in the ideological rhetoric of Democratic candidates. In addition, Republican candidates are consistently estimated to be more extreme than their Democratic opponents. Taken together with evidence that Republican partisans have become extreme, these results suggest that Republicans likely face stronger incentives to employ extreme rhetoric in primaries and consequently to moderate their rhetoric in generals.

The data show mixed evidence for heterogeneity in the extent of moderation among Republicans based on general election competitiveness. While the MNIR measure shows moderation to vary significantly with general election competitiveness, the other two measures show no such effect. I observe no significant differences among Democrats on any of the three measures. Moreover, I do



not find evidence that there is heterogeneity in the extent of moderation based on primary election competitiveness or incumbency status for either party.

As my data only capture a single election year, I am not able to fully separate strategic moderation from other global factors that might have affected rhetoric. Moreover, the single election cycle limits the sample size and statistical precision, particularly for the event study and heterogeneity analyses. Data from future election cycles could further strengthen the current analyses.

My results suggest several additional avenues for future research. First, increasing the richness of the candidate speech features would be advantageous. In particular, analyzing tweet attachments, such as photos, videos, or linked articles, could enhance the signal within individual observations. Alternatively, including other social media platforms, like Facebook, could allow for a finer grained temporal analysis.

Second, obtaining data on the competitiveness of an election in real time – particularly among primaries – would be beneficial. One of the main difficulties in identifying the timing of candidates' ideological shifts proves to be the differing dynamics and environments of the individual races. For example, a candidate may begin moderating before the primary election ends because they are certain that they have the primary locked up. While my primary competitiveness heterogeneity analysis attempted to address this issue, the measure only uses ex-post primary information and the sample lacks statistical power.

Third, while my results suggest that political candidates moderate over the course of the election, the normative implications of this movement remain to be determined. Future research should measure whether ideological moderation increases general election success probabilities, as well as whether more extreme candidates find more success in primary elections. In addition, important ethical questions exist as to whether candidates should practice what amounts to deception of their potential voters, as such ideological flexibility masks a candidate's true beliefs.

## REFERENCES

- Acree, Brice DL, Justin H Gross, Noah A Smith, Yanchuan Sim, and Amber E Boyd-stun.** 2020. "Etch-a-Sketching: Evaluating the post-primary rhetorical moderation hypothesis." *American Politics Research*, 48(1): 99–131.
- Agranov, Marina.** 2016. "Flip-flopping, primary visibility, and the selection of candidates." *American Economic Journal: Microeconomics*, 8(2): 61–85.
- Ansolabehere, Stephen, and Philip Edward Jones.** 2010. "Constituents' responses to congressional roll-call voting." *American Journal of Political Science*, 54(3): 583–597.
- Ansolabehere, Stephen, James M Snyder Jr, and Charles Stewart III.** 2001. "Candidate positioning in US House elections." *American Journal of Political Science*, 136–159.
- Bonica, Adam.** 2013. "Ideology and interests in the political marketplace." *American Journal of Political Science*, 57(2): 294–311.
- Bonica, Adam.** 2014. "Mapping the ideological marketplace." *American Journal of Political Science*, 58(2): 367–386.
- Brady, David W, Hahrie Han, and Jeremy C Pope.** 2007. "Primary elections and candidate ideology: Out of step with the primary electorate?" *Legislative Studies Quarterly*, 32(1): 79–105.
- Buccoliero, Luca, Elena Bellio, Giulia Crestini, and Alessandra Arkoudas.** 2020. "Twitter and politics: Evidence from the US presidential elections 2016." *Journal of Marketing Communications*, 26(1): 88–114.
- Burden, Barry C.** 2001. "The polarizing effects of congressional primaries." In *Congressional primaries and the politics of representation.*, ed. P.F. Galderisi, M. Ezra and M. Lyons, 95–115. New York:Rowman & Littlefield Publishers.
- Burden, Barry C.** 2004. "Candidate positioning in US congressional elections." *British Journal of Political Science*, 34(2): 211–227.
- Calvert, Randall L.** 1985. "The value of biased information: A rational choice model of political advice." *The Journal of Politics*, 47(2): 530–555.
- Chefer, Hila, Shir Gur, and Lior Wolf.** 2021. "Transformer interpretability beyond attention visualization." 782–791.
- Cox, Gary W.** 1990. "Centripetal and centrifugal incentives in electoral systems." *American Journal of Political Science*, 903–935.
- Croco, Sarah E.** 2016. "The flipside of flip-flopping: Leader inconsistency, citizen preferences, and the war in Iraq." *Foreign Policy Analysis*, 12(3): 237–257.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.** 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*.
- Downs, Anthony.** 1957. "An economic theory of political action in a democracy." *Journal of political economy*, 65(2): 135–150.

- Duck-Mayr, JBrandon, and Jacob Montgomery.** 2021. “Ends Against the Middle: Measuring Latent Traits When Opposites Respond the Same Way for Antithetical Reasons.” *Political Analysis*, Accepted pending replication.
- Enke, Benjamin.** 2020. “Moral values and voting.” *Journal of Political Economy*, 128(10): 3679–3729.
- Enke, Benjamin, Ricardo Rodriguez-Padilla, and Florian Zimmermann.** 2021. “Moral Universalism: Measurement and Economic Relevance.” *Management Science*.
- Fiorina, Morris P, and Matthew S Levendusky.** 2006. “Disconnected: The political class versus the people.” In *Red and blue nation.*, ed. Pietro S. Nivola and David W. Brady, 49–71. Washington, DC:Brookings Institution Press.
- Fiorina, Morris P, Samuel J Abrams, and Jeremy C Pope.** 2010. *Culture war: The myth of a polarized America.* New York:Longman.
- Gentzkow, Matthew.** 2016. “Polarization in 2016.” *Toulouse Network for Information Technology Whitepaper*, 1–23.
- Gentzkow, Matthew, and Jesse M Shapiro.** 2010. “What drives media slant? Evidence from US daily newspapers.” *Econometrica*, 78(1): 35–71.
- Gentzkow, Matthew, Bryan Kelly, and Matt Taddy.** 2019. “Text as data.” *Journal of Economic Literature*, 57(3): 535–74.
- Gentzkow, Matthew, Jesse M Shapiro, and Matt Taddy.** 2019. “Measuring group differences in high-dimensional choices: method and application to congressional speech.” *Econometrica*, 87(4): 1307–1340.
- Graham, Jesse, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto.** 2011. “Mapping the moral domain.” *Journal of personality and social psychology*, 101(2): 366.
- Graham, Jesse, Jonathan Haidt, and Brian A Nosek.** 2009. “Liberals and conservatives rely on different sets of moral foundations.” *Journal of personality and social psychology*, 96(5): 1029.
- Grossmann, Matt, and David A Hopkins.** 2015. “Ideological Republicans and group interest Democrats: The asymmetry of American party politics.” *Perspectives on Politics*, 13(1): 119–139.
- Guisinger, Alexandra.** 2009. “Determining trade policy: Do voters hold politicians accountable?” *International Organization*, 63(3): 533–557.
- Haidt, Jonathan, and Craig Joseph.** 2004. “Intuitive ethics: How innately prepared intuitions generate culturally variable virtues.” *Daedalus*, 133(4): 55–66.
- Haidt, Jonathan, and Jesse Graham.** 2007. “When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize.” *Social Justice Research*, 20(1): 98–116.
- Hall, Andrew B.** 2015. “What happens when extremists win primaries?” *American Political Science Review*, 109(1): 18–42.
- Hall, Andrew B, and James M Snyder.** 2015. “Candidate ideology and electoral success.” *Unpublished Manuscript, Stanford University.*

- Hannikainen, Ivar R, Ryan M Miller, and Fiery A Cushman.** 2017. "Act versus impact: Conservatives and liberals exhibit different structural emphases in moral judgment." *Ratio*, 30(4): 462–493.
- Hayes, Danny, and Jennifer L Lawless.** 2018. "The decline of local news and its effects: New evidence from longitudinal data." *The Journal of Politics*, 80(1): 332–336.
- Heaney, Michael T, Seth E Masket, Joanne M Miller, and Dara Z Strolovitch.** 2012. "Polarized networks: The organizational affiliations of national party convention delegates." *American Behavioral Scientist*, 56(12): 1654–1676.
- Hofmann, Wilhelm, Daniel C Wisneski, Mark J Brandt, and Linda J Skitka.** 2014. "Morality in everyday life." *Science*, 345(6202): 1340–1343.
- Hotelling, Harold.** 1929. "Stability in Competition." *The Economic Journal*, 39(153): 41–57.
- Hummel, Patrick.** 2010. "Flip-flopping from primaries to general elections." *Journal of Public Economics*, 94(11-12): 1020–1027.
- Iyengar, Shanto, and Adam F Simon.** 2000. "New perspectives and evidence on political communication and campaign effects." *Annual review of psychology*, 51(1): 149–169.
- Iyer, Ravi, Spassena Koleva, Jesse Graham, Peter Ditto, and Jonathan Haidt.** 2012. "Understanding libertarian morality: The psychological dispositions of self-identified libertarians." *PLoS ONE*, 7(8): e42366.
- Jacobson, Gary C, and Jamie L Carson.** 2019. *The politics of congressional elections*. Rowman & Littlefield.
- Jones, Philip Edward.** 2011. "Which buck stops here? Accountability for policy positions and policy outcomes in congress." *The Journal of Politics*, 73(3): 764–782.
- Jungherr, Andreas.** 2016. "Twitter use in election campaigns: A systematic literature review." *Journal of information technology & politics*, 13(1): 72–91.
- Kreiss, Daniel, and Shannon C McGregor.** 2018. "Technology firms shape political communication: The work of Microsoft, Facebook, Twitter, and Google with campaigns during the 2016 US presidential cycle." *Political Communication*, 35(2): 155–177.
- Lewis, Jeffrey B.** 2022. "Why is Alexandria Ocasio-Cortez estimated to be a moderate by NOMINATE?" [https://voteview.com/articles/ocasio\\_cortez](https://voteview.com/articles/ocasio_cortez).
- Lewis, Jeffrey B., Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet.** 2022. "Voteview: Congressional Roll-Call Votes Database." <https://voteview.com/>.
- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.** 2019. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692*.
- Luguri, Jamie B, Jaime L Napier, and John F Dovidio.** 2012. "Reconstruing intolerance: Abstract thinking reduces conservatives' prejudice against nonnormative groups." *Psychological science*, 23(7): 756–763.

- Martin, Gregory J, and Ali Yurukoglu.** 2017. "Bias in cable news: Persuasion and polarization." *American Economic Review*, 107(9): 2565–99.
- McCabe, David.** 2015. "Welcome to the social media election." <https://thehill.com/policy/technology/251185-welcome-to-the-social-media-election/>.
- McCarty, Nolan, Keith T Poole, and Howard Rosenthal.** 2016. *Polarized America: The dance of ideology and unequal riches*. Cambridge, MA:MIT Press.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean.** 2013. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781*.
- Pew.** 2014. "Political Polarization and Growing Ideological Consistency." <https://www.pewresearch.org/politics/2014/06/12/section-1-growing-ideological-consistency/#interactive>.
- Pew.** 2019. "How Twitter Users Compare to the General Public." <https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/>.
- Plott, Charles R.** 1967. "A notion of equilibrium and its possibility under majority rule." *The American Economic Review*, 57(4): 787–806.
- Poole, Keith T, and Howard Rosenthal.** 2017. *Ideology & congress: A political economic history of roll call voting*. Routledge.
- Porter, Martin F.** 1980. "An algorithm for suffix stripping. Program, 14 (3), 130-137." *N. Zhang and B. Ma/Constructing a Methodology Toward Policy Analysts*, 79.
- Rogowski, Jon C.** 2013. "Primary Systems, Candidate Platforms, and Ideological Extremity." *Washington University in St. Louis. Unpublished manuscript*.
- Smith, Isaac H, Karl Aquino, Spassena Koleva, and Jesse Graham.** 2014. "The moral ties that bind... even to out-groups: The interactive effect of moral identity and the binding moral foundations." *Psychological science*, 25(8): 1554–1562.
- Taddy, Matt.** 2013. "Multinomial inverse regression for text analysis." *Journal of the American Statistical Association*, 108(503): 755–770.
- Taddy, Matt.** 2015. "Distributed multinomial regression." *The Annals of Applied Statistics*, 9(3): 1394–1414.
- Tomz, Michael, and Robert P Van Houweling.** 2009. "The electoral implications of candidate ambiguity." *American Political Science Review*, 103(1): 83–98.
- Tomz, Michael, and Robert P Van Houweling.** 2014. "Political repositioning: A conjoint analysis." *Unpublished manuscript, Stanford University*.
- Tucker, Joshua A, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan.** 2018. "Social media, political polarization, and political disinformation: A review of the scientific literature." *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.** 2017. "Attention is all you need." *Advances in neural information processing systems*, 30.
- Wittman, Donald.** 1977. "Candidates with policy preferences: A dynamic model." *Journal of economic Theory*, 14(1): 180–189.
- Yaqub, Ussama, Soon Ae Chun, Vijayalakshmi Atluri, and Jaideep Vaidya.** 2017. "Analysis of political discourse on twitter in the context of the 2016 US presidential elections." *Government Information Quarterly*, 34(4): 613–626.

## APPENDIX

A1. *Most Extreme Senators*

TABLE A1—MOST EXTREME SENATORS, MNIR

	Name	Party	State	Predictions	True DW1	DW1 Rank
1	HARRIS, Kamala	D	California	-0.52	-0.71	2
2	SANDERS, Bernard	D	Vermont	-0.51	-0.53	4
3	BOOKER, Cory	D	New Jersey	-0.47	-0.61	3
4	WARREN, Elizabeth	D	Massachusetts	-0.47	-0.77	1
5	MURPHY, Christopher	D	Connecticut	-0.45	-0.28	29
6	MERKLEY, Jeff	D	Oregon	-0.45	-0.44	11
7	HIRONO, Mazie	D	Hawaii	-0.45	-0.51	6
8	SCHUMER, Charles	D	New York	-0.44	-0.36	19
9	BLUMENTHAL, Richard	D	Connecticut	-0.40	-0.43	12
10	MURRAY, Patty	D	Washington	-0.39	-0.35	21
...						
88	BLACKBURN, Marsha	R	Tennessee	0.53	0.62	83
89	CRUZ, Ted	R	Texas	0.58	0.82	89
90	SASSE, Benjamin	R	Nebraska	0.58	0.80	87
91	LEE, Mike	R	Utah	0.58	0.91	91
92	CASSIDY, Bill	R	Louisiana	0.59	0.45	66
93	INHOFE, James	R	Oklahoma	0.59	0.56	75
94	PAUL, Rand	R	Kentucky	0.59	0.88	90
95	CORNYN, John	R	Texas	0.59	0.49	70
96	BRAUN, Michael	R	Indiana	0.59	0.80	88
97	PERDUE, David	R	Georgia	0.59	0.57	78

*Notes:* This table presents the ten most extreme Republican and Democratic Senators according to the MNIR out-of-sample predictions using gradient boosting. The true DW-Nominate 1 (represented in the table as DW1) score and the corresponding rank is also included for reference. Due to Senators having equal true scores, the maximum “rank” is 91.

TABLE A2—MOST EXTREME SENATORS, ROBERTA

	Name	Party	State	Predictions	True DW1	DW1 Rank
1	MURRAY, Patty	D	Washington	-0.44	-0.35	21
2	HARRIS, Kamala	D	California	-0.41	-0.71	2
3	MARKEY, Edward	D	Massachusetts	-0.41	-0.51	5
4	WARREN, Elizabeth	D	Massachusetts	-0.40	-0.77	1
5	SANDERS, Bernard	D	Vermont	-0.40	-0.53	4
6	MERKLEY, Jeff	D	Oregon	-0.40	-0.44	11
7	UDALL, Thomas	D	New Mexico	-0.39	-0.45	9
8	BLUMENTHAL, Richard	D	Connecticut	-0.39	-0.43	12
9	MENENDEZ, Robert	D	New Jersey	-0.39	-0.37	17
10	SCHUMER, Charles	D	New York	-0.39	-0.36	19
...						
88	BOOZMAN, John	R	Arkansas	0.42	0.40	57
89	CRAPO, Michael	R	Idaho	0.44	0.51	71
90	COTTON, Tom	R	Arkansas	0.44	0.58	80
91	ROUNDS, Mike	R	South Dakota	0.44	0.40	56
92	ISAKSON, Johnny	R	Georgia	0.48	0.40	58
93	FISCHER, Debra	R	Nebraska	0.48	0.47	67
94	BLACKBURN, Marsha	R	Tennessee	0.50	0.62	83
95	CRAMER, Kevin	R	North Dakota	0.50	0.39	55
96	BRAUN, Michael	R	Indiana	0.51	0.80	88
97	CRUZ, Ted	R	Texas	0.51	0.82	89

*Notes:* This table presents the ten most extreme Republican and Democratic Senators according to the RoBERTa out-of-sample predictions. The true DW-Nominate 1 (represented in the table as DW1) score and the corresponding rank is also included for reference. Due to Senators having equal true scores, the maximum “rank” is 91.



### A2. Distribution of Fitted Candidate Predictions for MNIR

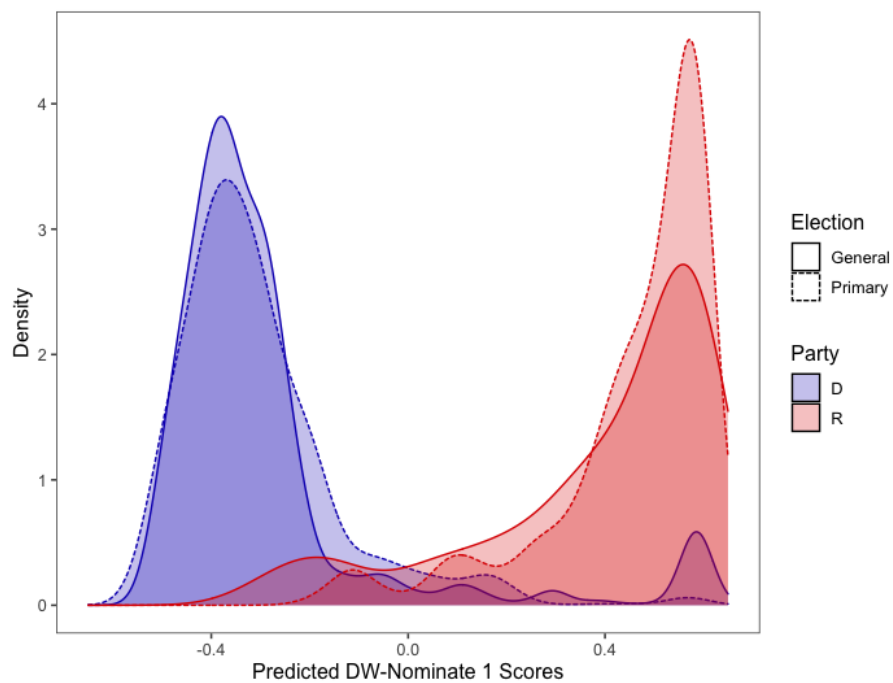


FIGURE A1. DISTRIBUTION OF FITTED CANDIDATE PREDICTIONS BY PARTY AND ELECTION

*Notes:* This figure visualizes the densities of the average fitted predictions from MNIR for candidates of both parties in the primary and the general. There is a visible moderating trend among Republicans but not Democrats.

### A3. Comparing Gradient Boosting and Regression Forest Results for MNIR

TABLE A3—IDEOLOGICAL MODERATION, BY MNIR FORWARD REGRESSION

	MNIR			
	Gradient Boosting		Regression Forest	
	(1)	(2)	(3)	(4)
Republican × General	−0.057** (0.022)	−0.058** (0.021)	−0.047*** (0.013)	−0.049*** (0.013)
Democrat × General	0.015 (0.038)	0.014 (0.036)	0.022 (0.029)	0.020 (0.027)
Republican	0.740*** (0.019)	0.745*** (0.019)	0.653*** (0.015)	0.657*** (0.015)
Trump 2016 Vote Share	0.339*** (0.082)	0.336*** (0.076)	0.397*** (0.079)	0.385*** (0.075)
Incumbent		−0.012 (0.021)		−0.022 (0.018)
Senate		−0.025 (0.024)		−0.035 (0.019)
Competitive		−0.044* (0.021)		−0.040* (0.017)
Constant	−0.455*** (0.037)	−0.433*** (0.032)	−0.431*** (0.039)	−0.400*** (0.038)
Observations			8,349	
Candidates			665	
Outcome Mean		−0.056		−0.031
Outcome SD		0.443		0.376
$R^2$	0.609	0.612	0.667	0.672
Hypothesis Tests ( $\chi^2$ )				
$\beta_{R,G} + \beta_{D,G} = 0$	−0.042 (0.045)	−0.044 (0.042)	−0.025 (0.033)	−0.029 (0.030)

*Notes:* This table presents the results from estimating Equation 2 on the final candidate sample for MNIR. The dependent variable is the predicted DW-Nominate scores from the gradient boosting forward regression for Columns (1) and (2); and from the regression forest for Columns (3) and (4). Odd-numbered columns control only for Trump's 2020 presidential vote share in the district; even-numbered columns also include indicators for incumbency status, competitiveness of the district (Cook PVI), and the congressional chamber. Observations are weighted by the number of bigram counts. All standard errors are clustered at the candidate level. In addition, the results of the hypothesis test on the equality of coefficients for the Republican and Democrat interaction terms are reported.

\* Significant at the 5% level. \*\* Significant at the 1% level. \*\*\* Significant at the 0.1% level.

*A4. Results Sensitivity to Sample Balancing*

TABLE A4—IDEOLOGICAL MODERATION, BY BALANCING THRESHOLD (MNIR)

	Sample Balancing			
	None (1)	50% (2)	75% (Default) (3)	Full (4)
Republican × General	−0.052* (0.020)	−0.049* (0.021)	−0.057** (0.022)	−0.052* (0.024)
Democrat × General	0.010 (0.036)	0.012 (0.036)	0.015 (0.038)	0.026 (0.041)
Republican	0.741*** (0.019)	0.740*** (0.019)	0.740*** (0.019)	0.740*** (0.021)
Trump 2020 Vote Share	0.320*** (0.075)	0.327*** (0.077)	0.339*** (0.082)	0.371*** (0.094)
Constant	−0.446*** (0.034)	−0.449*** (0.035)	−0.455*** (0.037)	−0.472*** (0.041)
Observations	9,448	9,108	8,349	7,097
$R^2$	0.611	0.611	0.609	0.61

TABLE A5—IDEOLOGICAL MODERATION, BY BALANCING THRESHOLD (ROBERTA)

	Sample Balancing			
	None (1)	50% (2)	75% (Default) (3)	Full (4)
Republican × General	−0.062*** (0.012)	−0.061*** (0.012)	−0.067*** (0.012)	−0.068*** (0.013)
Democrat × General	−0.030* (0.012)	−0.030* (0.013)	−0.029* (0.013)	−0.026 (0.014)
Republican	0.411*** (0.019)	0.410*** (0.019)	0.410*** (0.019)	0.413*** (0.021)
Trump 2020 Vote Share	0.585*** (0.060)	0.583*** (0.061)	0.583*** (0.065)	0.593*** (0.073)
Constant	−0.473*** (0.031)	−0.472*** (0.031)	−0.473*** (0.033)	−0.479*** (0.036)
Observations	9,423	9,088	8,304	7,046
$R^2$	0.575	0.576	0.576	0.585

TABLE A6—IDEOLOGICAL MODERATION, BY BALANCING THRESHOLD (MFD)

	Sample Balancing			
	None (1)	50% (2)	75% (Default) (3)	Full (4)
Republican $\times$ General	-0.084*** (0.019)	-0.082*** (0.018)	-0.096*** (0.019)	-0.093*** (0.018)
Democrat $\times$ General	-0.014 (0.015)	-0.020 (0.014)	-0.024 (0.014)	-0.022 (0.015)
Republican	0.275*** (0.019)	0.274*** (0.019)	0.269*** (0.019)	0.273*** (0.020)
Trump 2020 Vote Share	0.177** (0.057)	0.163** (0.057)	0.155* (0.062)	0.173** (0.065)
Constant	-0.167*** (0.030)	-0.160*** (0.031)	-0.156*** (0.033)	-0.164*** (0.034)
Observations	9,448	9,108	8,349	7,097
$R^2$	0.089	0.093	0.096	0.112

A5. *Results using MC3-GGUM adjusted DW-Nominate 1 Scores*

TABLE A7—IDEOLOGICAL MODERATION WITH MC3 - GGUM IDEOLOGY SCORES

	MNIR			
	Gradient Boosting		Regression Forest	
	(1)	(2)	(3)	(4)
Republican $\times$ General	-0.080*** (0.021)	-0.056*** (0.015)	-0.056*** (0.016)	-0.083*** (0.020)
Democrat $\times$ General	0.027 (0.044)	0.023 (0.030)	0.024 (0.031)	0.024 (0.042)
Republican	0.828*** (0.024)	0.678*** (0.018)	0.674*** (0.018)	0.832*** (0.024)
Trump 2020 Vote Share	0.420*** (0.106)	0.361*** (0.071)	0.357*** (0.077)	0.412*** (0.096)
Incumbent		-0.001 (0.019)		-0.026 (0.026)
Senate		-0.016 (0.023)		-0.060 (0.032)
Competitive		-0.041* (0.019)		-0.038 (0.025)
Constant	-0.583*** (0.054)	-0.487*** (0.033)	-0.499*** (0.037)	-0.548*** (0.049)
Observations			8,349	
Candidates			665	
Outcome Mean		-0.117		-0.111
Outcome SD		0.521		0.392
$R^2$	0.54	0.644	0.641	0.544
Hypothesis Tests ( $\chi^2$ )				
$\beta_{R,G} + \beta_{D,G} = 0$	-0.053 (0.050)	-0.033 (0.032)	-0.032 (0.036)	-0.059 (0.047)

This table presents the results from estimating Equation 2 on the final candidate sample for MNIR. The dependent variable is the predicted DW-Nominate scores using the MC3-GGUM adjustments from the gradient boosting forward regression for Columns (1) and (2); and from the regression forest for Columns (3) and (4). Odd-numbered columns control only for Trump's 2020 presidential vote share in the district; even-numbered columns also include indicators for incumbency status, competitiveness of the district (Cook PVI), and the congressional chamber. Observations are weighted by the number of bigram counts. All standard errors are clustered at the candidate level. In addition, the results of the hypothesis test on the equality of coefficients for the Republican and Democrat interaction terms are reported. The Chi-squared test statistic is reported.

\* Significant at the 5% level. \*\* Significant at the 1% level. \*\*\* Significant at the 0.1% level.

## A6. MNIR Text Processing

I originally noticed odd predictions for some candidates. For example, take Paula Jean Swearengin, the Democratic candidate for the West Virginia Senate Race. At first, MNIR classified her as conservative (with an average score of 0.37) while RoBERTa evaluated her rhetoric to be fairly left-leaning (with an average of -0.311). Most political pundits would agree with the RoBERTa classification, as she is considered to be quite progressive, endorsing Senator Bernie Sanders for the Democratic nomination and forming a primary challenge from the left to Senator Joe Manchin in 2018. So, why does MNIR classify her rhetoric as so conservative?

One likely explanation is that MNIR predicts mentions of West Virginia (“west\_virginians” or “west\_virginia”) to be right-leaning as it is a very conservative state and the majority of mentions in the baseline sample were from Republican representatives. Indeed, in the original baseline sample, the bigram “west\_virginia” is mentioned 44 times by a Democrat and 367 times by a Republican. Although Swearengin uses bigrams that are strongly on the left (“labor\_right” and “keep\_fighting”), this isn’t able to offset the frequency of Tweets referring to her home state. I re-estimate the fitted DW-Nominate scores for Swearengin throughout the cycle with the two West Virginia bigrams dropped. As can be seen in Figure A2, this has a dramatic effect, yielding a new average score of -0.463, which is significantly more left-leaning and in accord with expectations. A similar dynamic exists for North Dakota Senate candidate Zach Raknerud, who, due to the frequency of “north\_dakota” mentions is estimated to be extremely conservative with an average score of 0.33. Without these bigrams, Raknerud’s average fitted score is -0.28, which is in line with other moderate Democrats.

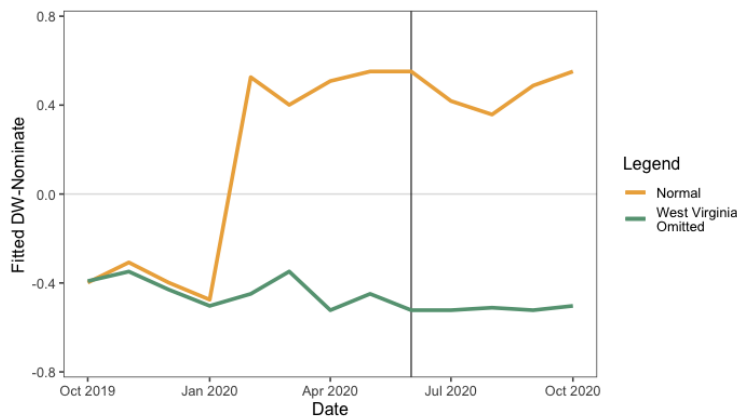


FIGURE A2. COMPARING FITTED DW-NOMINATE SCORES FOR SWEARENGIN, WITH AND WITHOUT WEST VIRGINIA BIGRAMS

As a result of this analysis, I returned to the preprocessing steps and removed all mentions of states and re-fit the model. Consequently, none of the selected bigrams can reference states, and thus the MNIR model cannot rely on the candidate’s home state to inform ideology prediction. The scores for the newly fit model closely resemble the scores of the old model with state bigrams omitted, as seen in Figure A2. For example, the average new score for Raknerud is -0.30, only 0.02 different. Hereafter, all mentions of the MNIR model will correspond to this newly fit model

without state references.