Estimating Ideology for U.S. House of Representatives Primary Candidates: Word Embeddings and Campaign Websites

Colin R. Case[*]

April 14, 2023

Abstract

While a broad range of research has focused on polarization and congressional primary elections (e.g. Thomsen 2014; Hall 2015; Bonica and Cox 2017), current measures of candidate ideology present significant limitations for the study of candidate behavior and polarization. To improve upon existing measures of congressional primary candidates' ideology, I use word embeddings with document-level vectors trained on congressional candidates' issue statements, as presented on their campaign websites. The estimates produced from this procedure are highly correlated with previous measures of elite ideology, such as DW-Nominate and CF Scores. I demonstrate the desirable properties of this measure, including (1) expanding the number of candidates with ideal-point estimates, (2) capturing candidate behavior versus perceptions of candidate behavior, and (3) the ability to validate the measure against underlying candidate behavior. Furthermore, I provide recommendations for researchers to consider when choosing a measure of candidate ideology.

[*] The University of North Carolina at Chapel Hill, Department of Political Science. Prepared for the Junior Americanist Workshop Series MPSA Mini-Conference on April 13, 2023. Please do not cite without permission.

As polarization in Congress has steadily risen since the 1970's, scholars have placed a substantive focus on understanding its causes and consequences. Congressional elections, and questions related to polarization and representation, have been central to this inquiry (e.g., Thomsen, 2014; Hall, 2015; Hall and Snyder, 2015). However, existing measures of candidate ideology limit the type of research questions that can be asked. For example, roll-call-based measures (e.g., Poole and Rosenthal, 1985) do not measure the behavior of non-incumbents and only focus on the legislative aspect of ideology (Bonica, 2014). Other approaches have expanded the measurement of ideology to candidates running for Congress. However, these measures still present limitations. For example, "expert" measures require significant time and resources to produce; even then, there is still disagreement about the placement of candidates (see Hirano et al., 2015). Other work estimating the ideology of congressional candidates has focused on the behavior of a subset of voters, such as Twitter followers (Barberá, 2015), donors (Bonica, 2014; Hall and Snyder, 2015), or surveys of voters (Christopher et al., 2015; Ramey, 2016). Across all these measures, there are two primary limitations. First, most measures of candidate ideology exclude significant populations of candidates, whether that be non-incumbents, inexperienced candidates, or candidates with a lower likelihood of success, but still have substantive importance (e.g., Porter and Treul, 2023; Treul and Hansen, 2023). For example, the most widely used measure of candidate ideology in primary elections, CFscores, did not have a score for 1,386 (35.5%) candidates for the U.S. House of Representatives in 2018 and 2020. Second, the vast majority of measures of candidate ideology base their estimation on *perceptions* of candidates versus actual candidate behavior. This data generating process presents significant validation and research limitations.

To improve upon these limitations of prior measures of candidate ideology, I propose a new measure of congressional candidate ideology using data collected from campaign website issue pages by Porter, Treul and Case (2023) for the 2018, 2020, and 2022 primaries for U.S. House of Representatives. Campaign websites are uniquely situated to study primary candidate behavior: they are unmediated in that they are directly from the campaign and not subject to other gatekeeping (e.g., media), complete over a range of policy areas, and represent all candidates running for the election, both experienced and inexperienced (Druckman, Kifer and Parkin, 2009). Further, campaign websites mitigate the limitations presented above by providing better coverage than donation-based measures (35.5% vs. 26.9%),[1] can compare scores with underlying campaign test, and capture actual candidate behavior versus perceptions of candidate behavior.

To estimate candidate ideology, I rely on recent developments in word embedding models that allow for the inclusion of a document-level vector for each candidate-year occurrence. This approach is validated across various contexts as a suitable way to uncover candidate ideology (Rheault and Cochrane, 2020). Further, because word embeddings uncover high-quality word embeddings in the same dimensional space as candidate embeddings, it is possible to carry out numerous validation procedures between the measurement and underlying text.

The paper proceeds as follows: I first outline existing measures of candidate ideology and their limitations. Turning to campaign websites, I demonstrate the extent to which data from campaign websites can improve upon current limitations with existing measures. I discuss the estimation strategy and conduct robustness checks to ensure ideology estimates are not sensitive to modeling decisions. I then validate the resulting measures against existing measures of candidate ideology covering a range of approaches. I also map word embeddings to the same dimensional space as candidates to demonstrate the ideology estimates are capturing ideological text in campaign website data. I conclude by providing recommendations to researchers using these measures in substantive research on candidate ideology and congressional elections a measurement for candidate ideology.

[1] As more candidates have issue pages on their website, this has improved to 20% in 2022

Measuring Ideology in Congressional Elections

Measuring candidates' preferences, or ideology, is crucial to the study of representation and elections in political science. Much of this work is grounded in the theory that voters prefer candidates proximal to their position (Black, 1948; Downs, 1957). Important substantive work has highlighted how voters can hold representatives accountable for these preferences, specifically related to voting behavior in Congress at the aggregate level (e.g., Canes-Wrone, Brady and Cogan, 2002) and on specific votes (e.g., Bussing et al., 2020). With the rise of polarization in Congress, research has turned to assess the ideological placement of members based on their roll call votes, with the most widely used measure being DW-Nominate (Poole and Rosenthal, 1985).[2] Other approaches have also used roll-call voting measures with different methodological approaches (Clinton, Jackman and Rivers, 2004) or assumptions (Duck-Mayr and Montgomery, 2022).

The lack of ideological measurement for non-incumbents presented research challenges for the study of congressional elections at the general and primary stages. In expanding the scope of study, some approaches link other measurements to predict unobserved legislative behavior (e.g. donor behavior in Hall and Snyder, 2015). The majority of measurement approaches turned to capturing candidate ideology without an empirical link to legislative behavior, but these approaches are often validated using DW-Nominate (Tausanovitch and Warshaw, 2017). While these measures provide little value in predicting future legislative behavior, especially within party (Tausanovitch and Warshaw, 2017), and should not be considered a measure of *legislative* ideology, candidate ideology is thought of as more than just how candidates may vote when in Congress. As such, *candidate* and *legislative* ideology should be considered related but conceptually distinct measurements. As Bonica (2014, pg. 372) notes:

[2] While there is an ongoing debate regarding the accuracy of DW-Nominate as a measure of ideology (e.g., Roberts, 2007; Lee, 2009; Duck-Mayr and Montgomery, 2022), it continues to serve as the preeminent measure by which scholars attempt to quantify members' preferences. Acknowledging the shortcomings of DW-Nominate's ability to do so, I will continue to refer to DW-Nominate, and other measures discussed below, as a measure of "ideology" in the absence of a more precise term.

> "Contributors are free to consider the many ways in which candidates express their ideology beyond how they vote, such as public-speaking records, stated policy goals, endorsements, the issues they champion, authored and cosponsored legislation, or cultural and religious values. As such, perfect correspondence between the two measures is neither expected nor necessarily desirable."

Further, given that voters are largely likely unaware of many of the votes members take in Congress outside of high-profile votes (Ansolabehere and Jones, 2010), candidate ideology represents a substantively important construct for studying the role ideology in congressional elections.

In the process of developing new measures of candidate ideology, a variety of approaches are taken. The most commonly used measures rely on the behavior of citizens, both through survey responses or actual behavior. In the various measurement approaches using the perceptions of candidates, different assumptions are imposed. However, all rely on the notion that the underlying data is capturing perceptions of candidates associated with stated policy goals and candidate ideology. Some approaches rely on voters' perceptions of candidates (Christopher et al., 2015; Ramey, 2016). In many ways, as those participating in elections, this approach is substantively meaningful. However, when comparing these results with the perception of experts, there are questions about citizens' ability to place candidates ideologically, especially in differentiating between candidates from the same party (Ahler, Citrin and Lenz, 2016). Further, knowledge and resource limitations make it challenging to produce widespread estimates of candidate ideology beyond general election candidates for Congress. For this reason, most estimates using voters' perceptions rely on the Cooperative Election Study (CES), but CES only asks about general election candidates. While expert surveys can mitigate concerns about citizens' ability to place candidates ideologically, these measures face the same resource constraints as other survey-based measures. Given primary elections for Congress are increasingly important with the decline of general election competition, measuring primary candidate ideology is a meaningful subset of candidates. Further, ideology in particular has been central to the renewed focus on primary elections (e.g. Thomsen, 2014; Hall, 2015; Hall and Snyder, 2015), making survey-based measures unsuitable for research questions of this type.

Other measurement approaches can circumvent these resource and knowledge constraints by focusing on the behavior of a subset of citizens, such as donors (Bonica, 2014) or Twitter users (Barberá, 2015). In these measurements, the estimation strategy assumes citizens' behavior is driven by a spatial model, either in which voters follow or contribute to. While most candidates running for Congress have a Twitter account, not all candidates receive donations. In 2018 and 2020, even as contributions have increased over time, 36% of candidates do not receive enough eligible contributions to have a valid CFscore. While most candidates who win recent general and primary elections receive donations (Thomsen, 2022), this is only sometimes the case. In 2018 and 2020, 155 candidates for the U.S. House of Representatives without a CFscore won their primary or were uncontested and appeared on the general election ballot. The 155 candidates on the general election ballot represent a sizable proportion of the general election contests where one candidate does not have an ideology score.

Further, not all research questions related to congressional elections focus on candidates who are more successful and more likely to have a CFscore. For example, in Congress, working-class Americans are descriptively under-represented, frequently due to obstacles related to their profession that make it difficult to run (Carnes, 2020). However, there is an open question of how these candidates actually perform when they do run. Prior work focused on if voters are biased against working-class candidates (Carnes and Lupu, 2016) or if these candidates struggle once reaching the general election stage (Carnes, 2020). Treul and Hansen (2023) provide more context to this question by evaluating the extent to which working-class candidates who run can win primary elections. After controlling for relevant factors, they find that working-class candidates are less likely to win and receive a lower vote share.

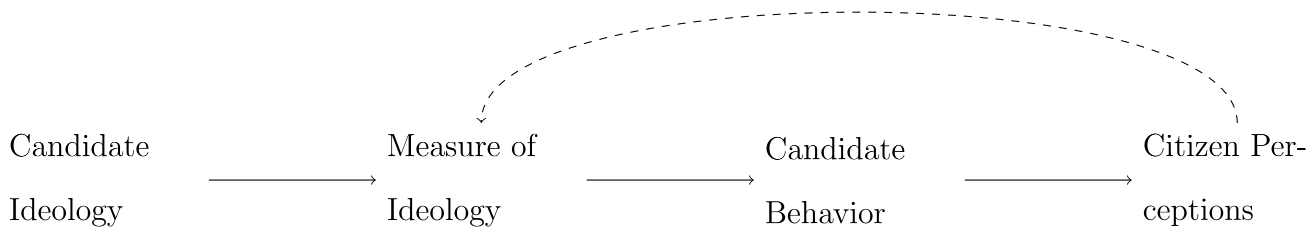
In the paper, the authors cannot control for ideology, a variable commonly used in a model for predicting candidate success. The authors note, "We omitted a measure of candidate ideology from the models above due to data limitations. Bonica's (2014) CFscores... are missing for 25% of all candidates in our data, including roughly 24% of nonworkers and 63%

of workers" (Treul and Hansen, 2023). While candidates who win primary elections will often have a CFscore due to the connection between fundraising and electoral success, there are substantively important research questions related to congressional elections that do not just focus on successful candidates. Treul and Hansen provide an illustrative example of where donation-based measures do not provide enough coverage of candidates with substantive importance.

Further, even if the behavior of general election candidates or incumbents is of substantive focus, understanding the ideology of primary challengers is still central to this inquiry. Members of Congress are responsive to the challengers they face and change their behavior as such (Sulkin, 2005). This plays out both in the legislative aspect of a members' job (Jewitt and Treul, 2014) and the issues candidates run on (Porter, McDonald and Treul, 2021).

In addition, while measures based on citizen perceptions often rely on spatial assumptions, the underlying data-generating process make it difficult to connect the measurement to changes in behavior. For example, in recent congressional elections, the intra-party correlation between CFscores and DW-Nominate diverges significantly (Barber, 2022). While Barber (2022) investigates a number of potential causes for this divergence, such as institutional, regional, and racial differences, as well as changes in donor behavior, the data-generating process underlying CFscores makes it difficult to adjudicate the reason; there is no underlying candidate behavior that can be connected back to validate changes in the measure.

Finally, ideology measures based on citizen perceptions are inappropriate for certain research questions. Consider the following DAG using a perception-based measure of candidate ideology. In the general theoretical framework where candidate ideology, measured by a perception-based measure, is associated with another candidate behavior, and that candidate behavior affects citizens' perceptions of candidate ideology, the measure of ideology becomes endogenous.



For example, Case (2023) argues that candidates use the visual elements of their campaign, such as the colors in their logo, to convey information to voters. Subsequently, these visual elements of a campaign affect voters' perceptions of candidates' ideology. In this instance, a measure based on perceptions of candidates would be endogenous; the link between the measure of candidate ideology and candidate behavior could be an artifact of the measurement.

Other measures avoid this by measuring ideology using candidate behavior instead of citizen perceptions. The most commonly used estimates of candidate ideology based on candidate behavior have focused on legislative positions outside of Congress. Given that historically the vast majority of non-incumbent candidates who are elected to Congress possess prior elected experience, this provides a good sample of potentially successful candidates. Most approaches rely on survey responses to Project Vote Smart's NPAT survey of legislators. For example, Shor and McCarty (2011) use each candidates' behavior in state legislatures separately and then link across different institutional bodies using the NPAT survey. Others, such as Ansolabehere, Snyder and Stewart (2001) and Montagnes and Rogowski (2015), rely solely on candidates' responses to the NPAT survey.

While both survey responses to the NPAT and behavior in state legislatures are the best predictor of freshman members' DW-Nominate score compared with other measures (Tausanovitch and Warshaw, 2017), there are limitations to using these measures in estimating candidate ideology. As with other survey-based measures, data limitations restrict ideology estimates to only candidates previously serving in state legislatures. While this estimate would have included a majority of freshman members in the past, recent trends have shown an increase in the number of incoming members without previous legislative experience (Porter and Treul, 2023). As such, this measure approach completely ignores a substantively important population of candidates. Further, while legislative ideology is likely related to the issues candidates run on, this measurement approach is not capturing campaign behavior per se.

Across all previously developed measures of candidate ideology, three issues are prevalent: first, the lack of coverage of candidates running in congressional elections, whether it be due to missing inexperienced candidates or candidates not running in the general election due to resource and knowledge limitations. Second, measures with high levels of coverage do not capture actual candidate behavior. This limits the ability to understand link changes in measurement validity with underlying candidate behavior. Further, these approaches can also be endogenous for specific research questions by relying on perceptions and not actual candidate behavior. In the following section, I discuss why issue positions on campaign websites are an ideal data source for overcoming these limitations in creating a measure of candidate ideology.

Data Description

Websites are an important part of each candidates' campaign. For most candidates in recent years (87.2% for 2018-2022), they maintain a website that acts as an "information hub" for all parts of the campaign, from information about the candidate to their issue positions and policy proposals (Herrnson, Panagopoulos and Bailey, 2019). Candidates carefully craft these websites, knowing that potential voters, donors, journalists, and other electoral stakeholders will visit them for information about the campaign (Druckman, Kifer and Parkin, 2009). Further, campaign websites come directly from the campaign, cover a range of issues and policy areas, and are representative of the population of campaigns (Druckman, Kifer and Parkin, 2009). Further, throughout an election cycle, little changes on the campaign website, making it a static representation of a campaign (Porter, McDonald and Treul, 2021). To this extent, campaign websites are a comprehensive data source to study candidates' behavior in

U.S. congressional elections.

As a part of their campaign website, most candidates maintain an "issue page" that explicitly lays out the candidates' opinions on the issues, specific policy proposals, and oftentimes commentary on contemporary events. Porter, Treul and Case (2023) collected the issue pages for all Democratic and Republican primary candidates for U.S. House of Representatives who had an official campaign website in 2018, 2020, and 2022. To collect official campaign website issue positions data, Porter, Treul and Case (2023) first identified the names of all candidates running in the primaries from state election boards as well as candidate filings with the Federal Election Commission (FEC). Using this list of names, official campaign websites were identified using a few different sources. Primarily, `Politics1.com` maintains a database of all campaign websites for candidates running actively in each race; this is where the links to majority of the campaign websites were found. Others were found through various social media pages, `Ballotpedia.com`, and through Google searches. Among 5,946 candidates who ran in for the House under either party label, Porter, Treul and Case (2023) found 5,188 (87.2%) candidates with campaign website. This represents a comprehensive data collection, both across years and in terms of coverage of candidates not just running in the general election.

As a part of this data collection process, research assistants identified whether or not each candidate had a "platform." While this looks different on some websites, it oftentimes was referred to as "My Platform," "Issues," or "Where I Stand." Interviews with campaign consultants who work with candidates on setting up their website highlighted the importance of these pages, mentioning issue pages as the part of the campaign they spent the most time discussing with candidates.[3] On these issue pages, candidates typically organize their issue

[3] It should be noted that while these campaign consultants often use similar strategies across campaigns (Nyhan and Montgomery, 2015), interviews highlighted a few important components that ensure the website is capturing candidate behavior. First, while campaign consultants help with the drafting process of issue pages, it is still what the candidate is interested in and wants to focus on for the election that shapes the issue pages. Second, candidates are still the ones operating their campaign, and even with the direction of campaign consultants, the candidate is the one with the final say. Third, despite consistent strategies across the same consulting firms, most have a review process across candidates to ensure that issue text for one candidate is not the same as issue text from another candidate at the same firm; most of the time this process involves separate writers for the issue pages and a secondary check of all issue text. In this manner, these issue pages are individual to each candidate.

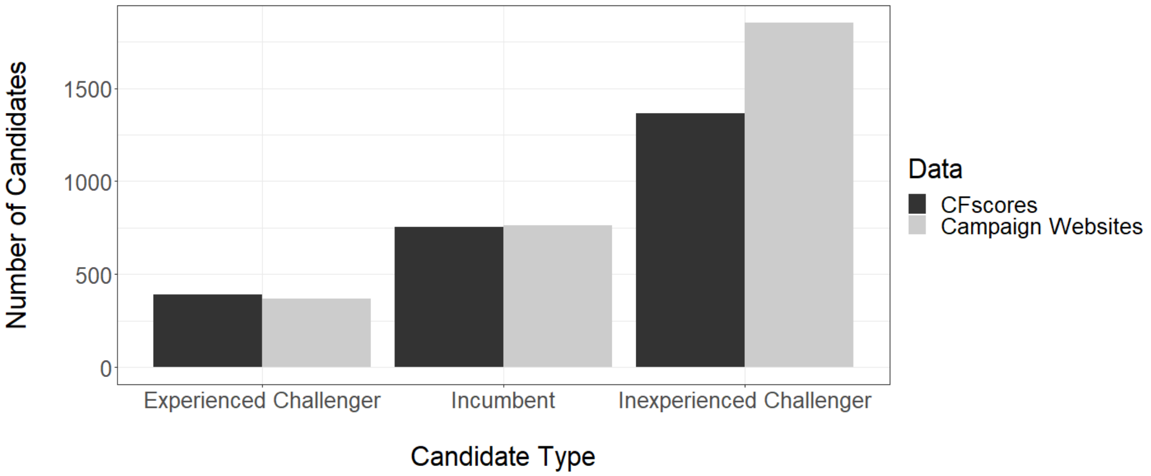
stances in a series of paragraphs about different policy areas, or individual issue statements. Research assistants manually collected each of these individual issue statements. This process was done in the ten days leading up to the primary to ensure consistency in candidates in the data collection process and that candidates websites had been finalized in the lead up to the election. As a part of this collection process, research assistants also identified the policy area for each issue statement into one of 23 issue areas.[4] In total, this data set contains 4,379 issue pages (73.6% of all candidates; 84.4% of candidates with a website). Across 2018, 2020, and 2022, each candidate had a mean of 9.7 for a total of 42,390 individual issue statements. For a full discussion of the data and the data collection process, see Porter, Treul and Case (2023).

Previous approaches to estimating candidate ideology have often focused on perceptions of candidates, in the hopes that these perceptions take into consideration the issues candidates run on, how they talk about specific issues, and what policies they propose. Rather than attempting to perceive candidates, campaign websites capture the actual expressed behaviors candidates convey that directly relate to candidate ideology. In this manner, campaign website issue positions represent a significant improvement in the link between concept and measurement, providing more face validity over prior measures. Given the measures are also based on actual text from the campaign, it is possible to validate the measurement with the underlying text that is closely associated with a candidate's ideology.

In addition, current measurements are not able to estimate ideology scores for a large proportion of candidates. While this coverage issue depends on the estimation strategy, most often, these coverage issues extend to inexperienced candidates, candidates with a lower likelihood of success, and those not on the general election ballot. While CFscores stand as measurement approach that includes primary candidates not on the general election ballot, they still exclude many of the other two types of candidates.

[4] The issue categories are: Abortion, Agriculture, Economy / Jobs, Education, Energy, Environment, Foreign Policy, Government, Group Issues (i.e. Women, LGBT, Civil Rights), Guns, Healthcare, Immigration, Infrastructure / Transportation, Local Issues, Military, Personal Characteristic (I am. . .), Political Opinions, Public Safety / Crime, Religion, Seniors, Advocacy for Vulnerable Pop., Support Troops / Veterans, Social Security, Unknown / Other

Figure 1: Candidate Coverage by Measurement and Candidate Type



To compare the coverage of candidates with an ideology measurement for CFscores and campaign websites, Figure 1 plots the number of candidates with an ideology score by previous political experience for the 2018 and 2020 U.S. House of Representatives primaries. This is further broken down by candidate type: incumbents, those who have previously held elected office, and those who have not previously held elected office. In the aggregate, 2,986 (77%) candidates had an issue page on their campaign website in 2018 and 2020 and 2,514 (65%) have a CFscore.

As is evident, campaign websites provide a significant increase in coverage of candidates when it comes to inexperienced challengers. Of the 2,652 inexperienced candidates who ran in 2018 and 2020, 1,854 (70%) had an issue page on their campaign website, while only 1,365 (51%) received enough eligible contributions for a CFscore. When it comes to experienced challengers, both sets of data have a high percentage of candidates, with 367 (76%) having an issue page and 392 (81%) out of 483 total experienced challengers having a CFscore in 2018 and 2020.[5] For incumbents, both measures have roughly 100% of candidates included.

Measurement coverage improves when focusing on candidates without a CFscore. In 2018 and 2020 there were 1,386 candidates without a CFscore. Among those, 745 (53.8%) have a campaign website issue page, representing a significant increase in coverage of these

[5] In 2022, these numbers were 225 (80%) for experienced challengers, 376 (100%) for incumbents, and 1,022 (74%) for inexperienced challengers.

candidates by using campaign websites. For research that is substantively interested in candidates who may not receive many donations, such as Treul and Hansen (e.g., 2023), or work interested in how primary challengers can shape member behavior (e.g., Jewitt and Treul, 2014; Porter, McDonald and Treul, 2021), this represents a substantial increase in the ability to test hypotheses related to ideology. Using underlying data that increases the number of candidates with a valid ideology estimate is an important advantage of campaign websites over other measures for ideological estimation.

Estimation Strategy

Estimating ideology from text has been an important application of substantive research, albeit a difficult task (Grimmer and Stewart, 2013). I rely on word embeddings to estimate ideology from candidates' campaign website text. Word embeddings are the parameter estimates from neural network models designed to predict word(s) given the context around that word(s). Work in other fields has highlighted the different ways in which word embeddings can capture important underlying properties of language, such as the similarity between words, analogies, and antonyms (Mikolov, Yih and Zweig, 2013).

Word embedding models have only recently seen more wide spread use in political science (Rodriguez and Spirling, 2022). Part of this stems from the ability to assess and test hypotheses for how word use can differ across covariates (Rodriguez, Spirling and Stewart, 2021) as well as uncover important latent traits related to the properties of both words (Grand et al., 2022) and the people using them (Rheault and Cochrane, 2020). Generally speaking, word embedding models represent words in a document as ordered sequences. In one of the more commonly used approaches, the continuous bag-of-words approach (CBOW) takes the Δ words before and after (referred to as the window) a target word, w_t, and uses those words as model inputs to predict the target words (Mikolov, Yih and Zweig, 2013). When it comes to the substantive application to political science, word embeddings can perform on a similar level as human coders on a Turing test explicitly designed for politically relevant terms (Rodriguez and Spirling, 2022).

There are a number of other approaches for measuring ideology in text. However, none of these other methods maintain the sequential ordering of text in the estimation procedure, missing important context in text related to ideology. Among the earliest approaches to estimating ideology using text is WordScores (Laver, Benoit and Garry, 2003). WordScores is a supervised machine learning method that uses a smaller sample of labeled documents, in the case of Laver, Benoit and Garry (2003) party manifestos, where each document has been labeled by experts to identify their ideological leaning. Based on the occurrence of each word in the labeled documents, words then receive a "score" representing their ideological lean. From there, unlabeled documents can then receive an ideological placement based on the occurrence of words and the scores for each word from the previous. WordFish (Slapin and Proksch, 2008) is another method that similarly relies on the occurrence of words in a document. Instead of using pre-labeled document, WordFish uses regressions to project counts for each word onto each party-year combination. More recently, Vafa, Naidu and Blei (2020) develop text-based ideal points (TBIP) that also uncover specific topics associated with each latent score, providing more validity and taking into account the co-occurrence of words.

While WordFish and TBIP improve upon supervised methods by reducing the time and cost of labeling documents, all three methods still rely on the occurrence (or co-occurrence in the case of TBIP) of words in a document without taking into account the full context of word usage. As Rheault and Cochrane (2020) note, while n-grams methods can include more context and focus on phrases with n-words rather than a single word, the estimation strategy becomes computationally inefficient. The focus on the occurrence of words rather than taking into account the full semantic context of a sentence presents substantive issues. Take, for example, the following two issue statements, one from a Democrat the other from a Republican, that discuss the Black Lives Matter movement:

> "Jamaal is fighting for an America where all of us have what we need to thrive.

> For an America where **Black lives matter**. For an America that finally belongs to all of us, not the wealthy few."
>
> "Dozens of Americans died during the extensive rioting during the summer of 2020. The groups perpetrating this violence have names. They are Antifa and **Black Lives Matter**. These militant groups are avowed Marxist revolutionaries who will stop at nothing to overthrow our Republican."

While Black Lives Matter is a phrase that is typically associated with more liberal candidates, a number of conservative candidates also use the phrase as well, leading to a u-shaped pattern associated with its use where both the most liberal candidates in the Democratic Party and most conservative candidates in Republican Party use the phrase compared with more moderate candidates (Bailey, 2023). Estimation approaches that focus solely on the occurrence of words can produce errors when these patterns emerge or need additional supervised steps that allow for subjective research decisions about the phrasing and context of a word. It should be noted TBIP get around this by focusing on the co-occurrence words. However, the more flexible model structure of neural networks is likely better suited to capture these complex semantic relationships related to ideology (Rheault and Cochrane, 2020).

Model Architecture

To estimate candidate ideology, I rely on a methodology similar to the one proposed by Rheault and Cochrane (2020) that has been validated in various political contexts to uncover valid estimates of elite ideology. This approach builds on traditional word embedding models by including indicator vectors for variables of interest at the document level (Xing and Jebara, 2014). In this approach, I use a shallow neural network between inputs and output data. The outcome variable, w_t, is the word occurring at position t in the text for all words in the vocabulary V. In the CBOW approach, the input variables are the Δ words that come before and after w_t in the corpus. This can be written more completely as $w_\Delta = (w_{t-\Delta}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+\Delta})$. In addition, I also include a document-level vector as an input, x, with size N for each candidate running in each election year.

The model estimation features two components. The first component estimates hidden nodes, z_m, for each node in the hidden layer with size M. Each node is subsequently represented as follows, where f is the activation function that takes the average across all inputs, and β_m and ζ_m are the embeddings for words and candidates respectively:

$$z_m = f(w'_\Delta \beta_m + x' \zeta_m)$$

Because w_Δ and x are indicator vectors, each hidden node can be expressed as

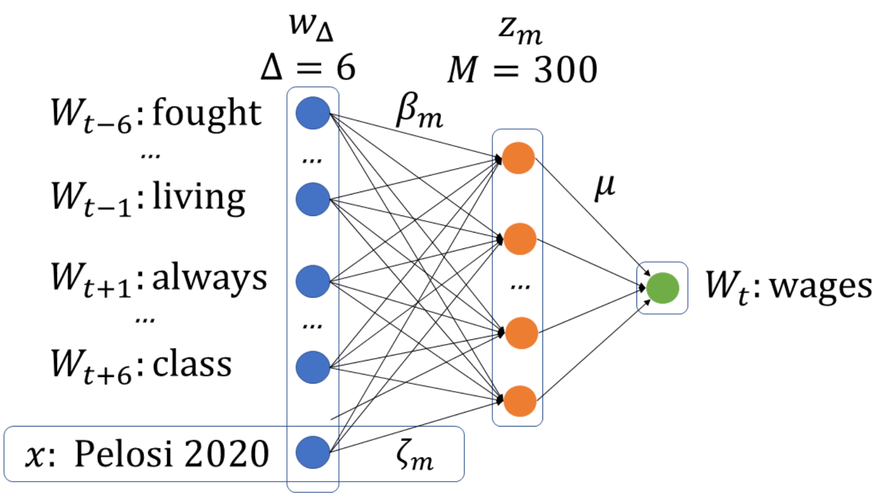
$$z_m = \frac{1}{2\Delta + 1}[(\sum_{w_v \in w_\Delta} \beta_{v,m}) + \zeta_m]$$

where $\sum_{w_v \in w_\Delta} \beta_{v,m}$ gives the sum of all coefficients $\beta_{v,m}$ for words v that exists in the window w_Δ for the dimension m. Similarly, ζ_m gives the coefficient for candidate i for the dimension m. Expanding this to the full vector of hidden nodes, z can be represented as

$$z = \frac{1}{2\Delta + 1}[(\sum_{w_v \in w_\Delta} \beta_v) + \zeta]$$

In estimating the model, the resulting β_v vector for each word in the vocabulary V and ζ vector for each candidate represents the word embeddings and document-level embeddings of interest. Figure 2 provides a graphical depiction of this process. The second part of the model is responsible for expressing the probability of the target word w_t as a function of the hidden layer z. To do this, a latent variable, u_{jt} for each word j can be expressed as a function of the hidden node, z.

Figure 2: Model Architecture with Window of 6 and Embedding Dimension of 300



In summary and application, the model is trained to predict the next word in a sentence given the (1) context surrounding the word and (2) the candidate. In traditional word embeddings, the target word is predicted by only the context around it. Take the issue statement examples below where healthcare is the target word (italicized) and the input words surrounding the target word (bolded). In traditional word embeddings, words such as "drug" and "quality" are likely informative for the model that healthcare is the target word. Further, it is easy to surmise how candidates across the ideological spectrum would use these words in close association. Thus a candidate embedding provides little value in the prediction of the word healthcare.

> "We pay 70% more in costs because of insurance companies, billing costs, hospital **administration, and drug companies. A national** *healthcare* **system has stronger buying power and** can negotiate lower prices for drugs and medical equipment as well as curb the astronomically high administrative salaries." - Rep. Alexandria Ocasio-Cortez
>
> "We need to continue to look for creative ideas that rely on the **marketplace to improve the quality of** *healthcare* **and empower individuals in their healthcare** decisions." - Speaker Kevin McCarthy

For other words, such as "national," or "marketplace," their usage in relation to healthcare is going to vary across the ideological spectrum: more liberal candidates will talk about a nationalized healthcare system while more conservative candidates will talk about marketplace competition in healthcare. By including a document level vector for each candidate, the candidate embedding is training on semantic differences between words in each candidates' issue positions. Subsequently, these differences can pick up on important changes in language that are related to what can conceptually be thought of as candidate ideology.

Model Implementation and Fit

In fitting the word embedding model on campaign website text, there are several decisions related to implementation, both pre-estimation, model parameters, and in creating the resulting measure, that need to be discussed. First, when pre-processing text, I remove all stop words. While word embedding models traditionally leave in stop words, I follow Rheault and Cochrane (2020) in focusing on the relationship between words with more substantive meaning. In addition, I also remove numbers and words with fewer than 5 occurrences to avoid overfitting.

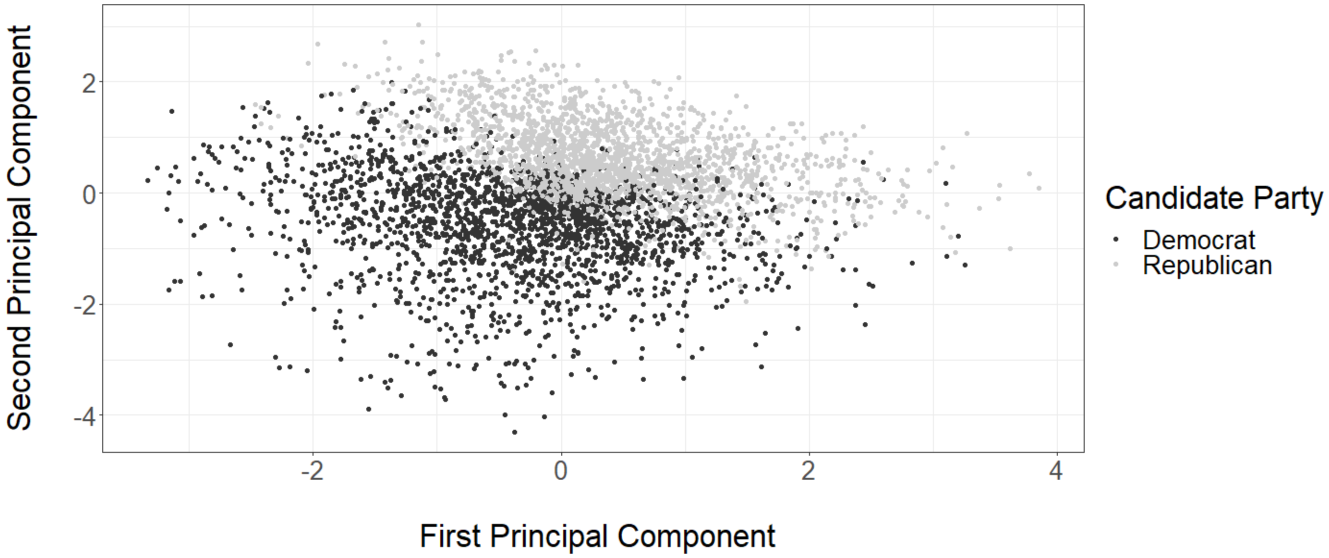
The model is fit using parameter recommendations from Rodriguez, Spirling and Stewart (2021), including a window of 6 and an embedding dimension of 300. Because there is no clear-cut justification for model parameters, I also fit models with various window sizes (5, 6, 7, and 8) and embedding dimensions (100, 200, 300) and show that resulting ideology measures are highly correlated (≥ 0.8). I also follow Mikolov, Yih and Zweig (2013) and use a learning rate of 0.025 and five epochs. Because ideology is commonly thought of on a liberal-conservative scale, I use principal component analysis to reduce the candidate embedding from 300 dimensions to two dimensions.

Analysis

In this section, I provide an overview of the measurement and several validation procedures. The resulting first two principle components of the candidate embeddings are plotted in

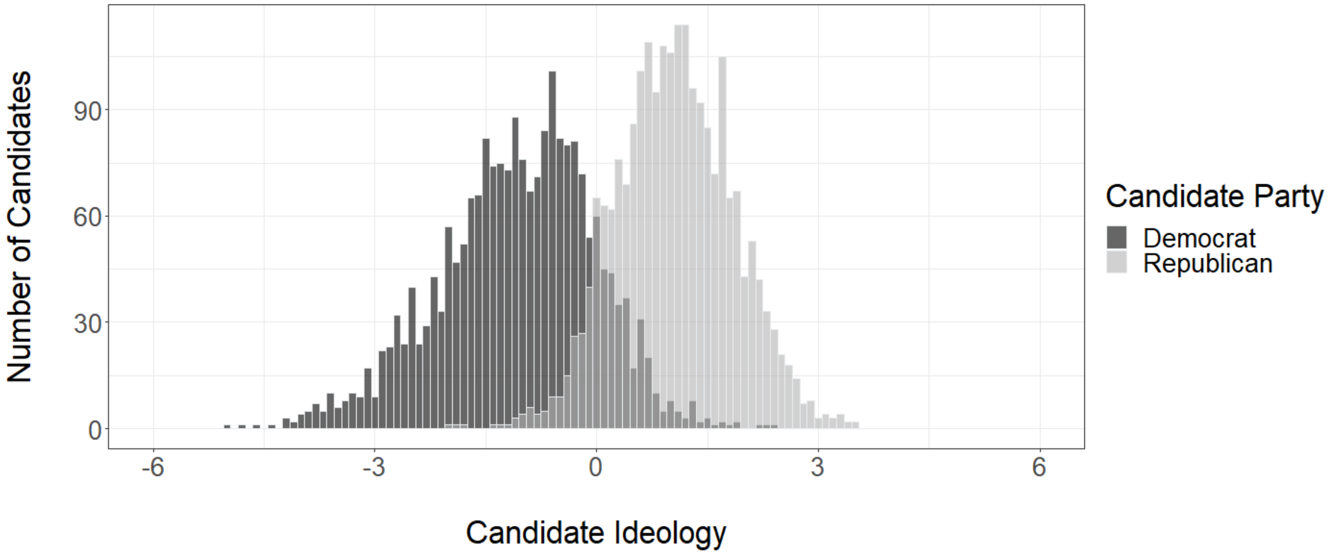
Figure 3. In general, the second dimension has more divergence between parties: the mean value on the first dimension for Democrats is -0.39 while the mean value for Republican is 0.39. On the second dimension, this is -0.60 and 0.59 for Democrats and Republicans respectively.

Figure 3: Scaled Two-Dimensional Projection of Candidate Embedding



Ideology along a single dimension is calculated by first standardizing the first two principle components then adding them. The resulting distribution of candidates is plotted in Figure 4. The measurement has a mean of zero across candidates and a standard deviation of 1.42. Democratic candidates trend to the negative side of the scale, as is standard convention, with a mean ideology score of -.99 while Republicans have a mean ideology score of 0.97.

Figure 4: Histogram of Candidate Ideology for Candidates in 2018, 2020, and 2022 U.S. House of Representatives Primary Election



To provide context to the measurement, Table 1 presents the ten most liberal and conservative incumbent Democratic candidates for U.S. House of Representatives from 2018-2022. Table 2 presents the corresponding table for Republicans. At face value, the measurement picks up on clear policy differences that map to candidate ideology and the positions candidates run on. Among the most liberal Democratic candidates, Rep. Jamaal Bowman ran on the Green New Deal, Medicare for All, and advocated for policies that would tackle income inequality. He also advocated for abolishing ICE, and creating a path to citizenship for all undocumented immigrants in the United States.

On the conservative side of the measurement, Rep. Bill Foster advocated for policies that stand in stark contrast. Foster has stayed a strong advocate for strengthening the Affordable Care Act through a public option that will foster competition while maintaining employer-sponsored healthcare. Further, while Foster supports the DREAM Act, he advocates for additional funding for border security and a “strict but fair” pathway to citizenship, conditional on paying fines, back taxes, and passing a criminal background check.

Further, when looking at the most liberal and conservative incumbents from the Democratic Party, assessing caucus membership can provide face validity to the results on the

Table 1: Most Liberal and Conservative Incumbent Democratic Candidates

Most Liberal	Most Conservative
Jamaal Bowman (2022)	Bill Foster (2018)
Rashida Tlaib (2022)	Bennie Thompson (2020)
Cori Bush (2022)	Brad Sherman (2020)
Ilhan Omar (2022)	Jerry McNerney (2018)
Carolyn Maloney (2022)	Jerry McNerney (2020)
Ilhan Omar (2020)	Bennie Thompson (2018)
Alexandria Ocasio-Cortez (2022)	Bill Foster (2020)
Ayanna Pressley (2020)	Ed Perlmutter (2020)
Carolyn Maloney (2020)	David Scott (2018)
Pramila Jayapal (2018)	David Scott (2020)

ideological leanings of members. Members view their caucus membership as a way to brand their ideological alignment within their parties to voters and donors (Clarke, 2020). Among the ten most liberal members by word embeddings ideology scores, all ten either are or were members of the more liberal Congressional Progressive Caucus. Among the more conservative Democratic members, Bill Foster, Ed Perlmutter, and David Scott were or are all members of the New Democrat Coalition and David Scott was a part of the Blue Dog Coalition prior to 2022, both of which are more moderate ideological caucuses in the Democratic Party.

On the Republican side, Table 2 also looks at the most liberal and conservative Republican incumbent candidates running in 2018-2022. As with Democrats, the differences in the issues candidates run on on each side of the spectrum provide face validity to the measurement. For Lee Zeldin in 2018, one of the more liberal Republican candidates, he relied strongly on talking about his bipartisan legislation. This included legislation improving lending to families and small businesses, rolling back mandatory educational testing, and combating the opioid epidemic. What Zeldin does not discuss also has substantive importance without a statement on the national debt, guns, or immigration in 2018.

On the conservative end of the Republican Party, McClintock is clear about his stance on the national debt, offering numerous proposals to ensure Congress does not increase government spending. He also discusses repealing and replacing the Affordable Care Act

Table 2: Most Liberal and Conservative Incumbent Republican Candidates

Most Liberal	Most Conservative
Nicole Malliotakis (2022)	Tom McClintock (2018)
Lee Zeldin (2018)	Rob Woodall (2018)
John Faso (2018)	Tom McClintock (2020)
Elise Stefanik (2018)	Mark Sanford (2018)
Brian Fitzpatrick (2022)	Chip Roy (2020)
Lee Zeldin (2020)	Chuck Fleishmann (2018)
Elise Stefanik (2022)	David Rouzer (2020)
Elise Stefanik (2020)	Steve Russell (2018)
John Rutherford (2022)	Tom McClintock (2022)
Rod Blum (2018)	Thomas Massie (2018)

with a plan that provides individuals with more choice and risk pools for those with pre-existing conditions. McClintock also advocated for an economic system with lower taxes and fewer regulations. The differences here between Zeldin and McClintock provide some level of face validity the measurement is picking up on differences within parties. Again, using ideological caucus membership as a test for face validity, the most conservative and liberal members in the Republican Party provide validation of the measurement. Among the most conservative members, Tom McClintock, Mark Sanford, Chip Roy, Chuck Fleishmann, and David Rouzer are or were all members of the more conservative Republican Study Committee. Roy and Sanford are also members of the House Freedom Caucus. On the moderate side, Nicole Malliotakis, Lee Zeldin, John Faso, Elise Stefanik, and John Rutherford either are or were members of the more ideologically moderate Main Street Partnership.

In the following sections, I carry out a variety of tests to evaluate the validity of the measurement. I first assess the external validity by comparing the measurement with previous measures of ideology (DW-Nominate and DIME) as well as validating the measurement consistency against candidates' issue positions.

External Validity

To assess the validity of the word embedding ideology measure, I first evaluate the word embeddings ideology score with pre-existing scores of ideology. For this, I focus on only 2018

Table 3: Word Embeddings Ideology and CFscore Correlations

	Correlation
All Candidates	0.70
Democrats	0.07
Republicans	0.20

and 2020 candidates, as the 118th Congress is still in session at the time of writing and the DIME database has been only updated to reflect 2018 and 2020 candidates. The correlations are broken down into two tables. Table 3 shows the correlations (both all candidates and intra-party) for candidates with both a CFscore and a word embedding score. Table 4 reflects the same comparisons for candidates who were elected to Congress. I also include correlations between CFscores and DW-Nominate as a baseline for comparison over the same period with the same set of candidates.

Starting with Table 3, correlations with all candidates are relatively high between CFscores and word embedding ideology scores at 0.70. It should be noted, one of the most significant differences between CFscores and word embeddings ideology is the amount of overlap between candidates from each party. According to CFscores, very few Democrats in 2018 and 2020 were more conservative than the mean; the same holds for Republicans and being more liberal than the mean. As is evident in Figure 4, word embedding ideology scores have more cross-party overlap. This is consistent with other text-based ideology measures Gaynor et al. (e.g. 2022). In a polarized era, intra-party correlations are important to validating a score (Tausanovitch and Warshaw, 2017). Intra-party correlations are weakly correlated at best, with a correlation of 0.07 for Democratic candidates and 0.20 for Republican candidates. In the aggregate, the results suggest that both CFscores and word embedding ideology scores are capturing distinct concepts within party, and further validation is needed.

Table 4 shows the correlations for candidates elected to Congress in 2018 and 2020, thus having a word embedding ideology score, a CFscore, and a DW-Nominate score for the 116th and 117th Congress. The first panel looks at all candidates, the second looks at Democratic candidates, and the third looks at Republican candidates. Starting with all candidates, the correlation between word embedding ideology scores and DW-Nominate is high at 0.79, albeit slightly lower than the correlation between DW-Nominate and CFscores at 0.92. It should also be noted that the correlation for candidates elected to Congress increases between CFscores and word embedding ideology scores, from 0.70 in Table 3 to 0.78. Part of this can be attributed to the fact that most incumbents receive a sizable number of donations that can be included in the CFscore calculation. When this occurs, CFscores and word embedding ideology scores become more closely related.[6]

Turning to intra-party correlations for Democratic candidates, word embedding ideology scores are weakly correlated with DW-Nominate at 0.16. [7] This is, however, higher than the correlation between CFscores and DW-Nominate for Democrats (-0.02). This trend in 2018 and 2020 is consistent with Barber (2022). In addition, as with the all candidate correlation, the correlation between CFscores and word embeddings increases when looking at just elected candidates versus all candidates (0.28 versus 0.07).

Among Republican candidates who were elected to Congress, the correlation between word embeddings and DW-Nominate is moderately higher than other intra-party correlations at 0.37. This is, however, slightly lower than the intra-party correlations for CFscores and DW-Nominate at 0.52. Further, unlike with Democrats, the intra-party correlation among elected Republican candidates is lower with word embeddings and CFscores at only 0.14. Overall, the correlations among elected officials between DW-Nominate, CFscores, and word embedding ideology scores all point to the measures capturing somewhat of a similar construct. However, there are significant divergences regarding candidates with fewer donations, and intra-party correlations within the Democratic Party specifically.

[6] For reference, when correlating CFscores and word embedding ideology scores with candidates who received 50 eligible donations or fewer, the all candidate correlation drops to only 0.55.

[7] It should be noted that part of this can likely be attributed to shortcomings of DW-Nominate in correctly identifying more liberal members of Congress, such as Alexandria Ocasio-Cortez, who often vote against legislation supported by Democrats because it is not liberal enough. Correlations with alternative estimation strategies, such as Duck-Mayr and Montgomery (2022), would likely be higher within the Democratic Party given this adjustment.

Table 4: Ideology Measure Correlations for 116th and 117th Congress

All Members of Congress			
	CFscores	DW-Nominate	Word Embeddings
CFscores	–	0.92	0.78
DW-Nominate	0.92	–	0.79
Word Embeddings	0.78	0.79	–

Democrats			
	CFscores	DW-Nominate	Word Embeddings
CFscores	–	-0.02	0.28
DW-Nominate	-0.02	–	0.16
Word Embeddings	0.28	0.16	–

Republicans			
	CFscores	DW-Nominate	Word Embeddings
CFscores	–	0.52	0.14
DW-Nominate	0.52	–	0.37
Word Embeddings	0.14	0.37	–

Internal Validity

One of the primary advantages of text-based ideological estimation, and specifically word embeddings, is the ability to validate data in the underlying text that is conceptually linked to ideology. On of the primary advantages of word embeddings with document level vectors is that both the candidate and word embeddings exist in the same dimensional space. Taking the same PCA transformation that converted candidate embeddings into the two dimensions, it is possible to locate words in the same dimensional space. This produces a substantive interpretation of various points on the two PCA dimensions (Rheault and Cochrane, 2020).

To do this, I follow the same process used by Rheault and Cochrane (2020). Each cardinal point on the two dimensions is determined by taking the maximum and minimum value of that dimension, and leaving the other dimension at zero. From there, I rank words by their euclidean distance to each of these cardinal points. I also do this for the southwest and northeast quadrant given the dimensional alignment presented in Figure 3. The top 20 closest words for each dimension are presented in Table 5.

As is evident, the different cardinal points on the axis are capturing policy relevant language that can be associated with various ideologies. Starting with the northern most

Table 5: Words and Phrases Closest fo PCA Axis

Quadrant (Dim 1, Dim 2)	Top Words
North (0, +)	hardearned, rick, fiscally, outofcontrol, thompson, producers, tape, jobkilling, liz, wasteful, progrowth, red, murphy, barack, scott, vigilant, rein, pentagon, iraq, redtape.
South (0, −)	indigenous, disproportionate, tenants, accommodations, restorative, excluded, bias, injustices, cycles, segregation, disparity, nonbinary, racial, unequal, exacerbated, impacts, solitary, marginalized, poorest, deescalation
East (+, 0)	socialism, govern, bible, winners, evil, whatsoever, militia, sounds, seem, isnt, probably, venezuela, fetus, document, experiment, losers, founders, dont, location, argue
West (−, 0)	cochair, pramila, alzheimer, reauthorization, caucus, maternal, rep, rosa, champion, supporter, anthony, espaillat, landmark, efforts, pell, shalala, she, coleman, affordability, hyde
Southwest (−,−)	frontline, lowincome, nondiscrimination, disproportionately, color, lgbtqia, antidiscrimination, gap, africanamerican, schooltoprison, equity, workplace, cleanup, lgbt, inequities, hiv, prek, evictions, workplaces, atrisk
Northeast (+,+)	lord, founders, creator, infringe, ronald, bureaucrats, fits, militia, jefferson, currency, statue, framers, putin, socialized, tyranny, swamp, tariffs, jesus, god, winners

point (conservative) on the two-dimensional scale, many of the terms relate to government spending, such as “out-of-control,” “fiscally,” and “rein,” as well as general government regulations with words such as “redtape.” Terms on the eastern most point (conservative), such as “bible,” “militia,” and “fetus,” also reflect what can be deemed more conservative policy positions. On the southern most point (liberal), words closest to this cardinal points are capturing liberal positions on social issues, such as “indigenous,” “injustices,” and “restorative.” The western most point (liberal) is more mixed and seems to capture language focused much more on legislative accomplishments with words such as “caucus,” “champion,” and “supporter.” There are still some policy related words such as “alzheimer” and “affordability.”

I also include the words closest to the cardinal points that represent the southwest (liberal) and northeast (conservative) points on the two dimensional plot. Given ideology scores are calculated by adding the two dimensions, these points in space are near the ideological endpoints of the unidimensional scale. Just like the cardinal points, both the southwest and northeast portion of the two-dimensional plot represent ends of the liberal conservative spectrum. Words such as “lowincome,” “nondiscrimination,” and “pre-k” all are closest to the liberal end of the scale while words such as “founders,” “bureaucrats,” and “tariffs” all are closest to the conservative end of the scale. The closest words provide substantive validation that the axis is picking up on ideological differences.

Conclusion

This paper presents an alternative measure of candidate ideology that is correlated with previous measures and reflects ideological differences in campaign website text. In addition, the underlying data increases the number of candidates with an ideology score. As discussed above, there are a wide range of measures of ideology, all with different assumptions, underlying data, and coverage of candidates. In this section, I offer advise to researchers interested in candidate ideology, either as a focal point of study or a model control, on what to consider when choosing a measure of ideology.

As a starting point, scholars interested in candidate ideology should ensure their theoretical quantity of interest is candidate ideology. An important limitation of word embedding ideology scores, and other measures as well is that they provide little in terms of predicting future legislative behavior. Rather, these measures are intended to capture the issues candidates run on, and how that reflects ideologically. Depending on the research question, word embedding ideology scores may not be appropriate. See Tausanovitch and Warshaw (2017) for a discussion of measures relation to future legislative behavior.

Second, researchers need to take into consideration the population of candidates they are interested in. Compared with previous measures, word embedding ideology scores provide more candidate coverage, especially for inexperienced candidates and candidates in primary elections. There is also reason to believe these estimates may be more precise than other measures like CFscores that rely on a small subset of candidates of interest. One explicit limitation of word embedding ideology scores is the broader scope of the measure; while Porter, Treul and Case (2023) have collected data for 2018, 2020, and 2022, the data collection does not exist prior to 2018. Research with a larger time frame is better suited to pick a measure such as CFscores. In addition, word embedding ideology scores cannot be calculated for potential candidates who may run for Congress. If researchers are interested in studying candidate emergence, both recipient CFscores and contributor CFscores provide more leverage of studying those who do not actually end up running for office.

Third, and finally, researchers need to consider their theoretical argument when choosing a measure of candidate ideology. There are a range of research questions where, if the measure approach is based on the perception, the measure could be endogenous. In these instances, it may be the case that citizens' perceptions are not only based on the candidates' actual ideology, but other related behaviors of interest. When this occurs, it is important to choose a measure that is capturing actual candidate behavior, like word embedding ideology scores, instead of a perception-based measure.

References

Ahler, Douglas J, Jack Citrin and Gabriel S Lenz. 2016. "Do Open Primaries Improve Representation? An Experimental Test of California's 2012 Top-Two Primary." *Legislative Studies Quarterly* 41(2):237–268.

Ansolabehere, Stephen, James M. Snyder and Charles Stewart. 2001. "Candidate Positioning in U.S. House Elections." *American Journal of Political Science* 45(1).

Ansolabehere, Stephen and Philip Edward Jones. 2010. "Constituents' Responses to Congressional Roll-Call Voting." *American Journal of Political Science* 54(3):583–597.

Bailey, Michael. 2023. "Measuring Candidate Ideology from Congressional Tweets and Websites." *Available at SSRN 4350550* .

Barber, Michael. 2022. "Comparing campaign finance and vote-based measures of ideology." *The Journal of Politics* 84(1):613–619.

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political analysis* 23(1):76–91.

Black, Duncan. 1948. "On the rationale of group decision-making." *Journal of political economy* 56(1):23–34.

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

Bussing, Austin, Will Patton, Jason M Roberts and Sarah A Treul. 2020. "The electoral consequences of roll call voting: Health care and the 2018 election." *Political Behavior* pp. 1–21.

Canes-Wrone, Brandice, David W. Brady and John F Cogan. 2002. "Out of step, out of Office: Electoral Accountability and House members' voting." *Annual Review of Political Science* 96:127–140.

Carnes, Nicholas. 2020. *The Cash Ceiling: Why Only the Rich Run for Office–and What We Can Do about It.* Princeton University Press.

Carnes, Nicholas and Noam Lupu. 2016. "Do voters dislike working-class candidates? Voter biases and the descriptive underrepresentation of the working class." *American Political Science Review* 110(4):832–844.

Case, Colin R. 2023. "Coloring within the Party Lines: Candidate Branding in Primary Elections.".

Christopher, Hare, Bakker Ryan, Carroll Royce et al. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3):759–774.

Clarke, Andrew J. 2020. "Party Sub-Brands and American Party Factions." *American Journal of Political Science* 64(3):452–470.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The statistical analysis of roll call data." *American Political Science Review* 98(2):355–370.

Downs, Anthony. 1957. "An Economic Theory of Political Action in a Democracy." *Journal of Political Economy* 65:135–150.

Druckman, James N., Martin J. Kifer and Michael Parkin. 2009. "Campaign Communications in U.S. Congressional Elections." *American Political Science Review* 103(3):343–366.

Duck-Mayr, JBrandon and Jacob Montgomery. 2022. "Ends against the middle: Measuring latent traits when opposites respond the same way for antithetical reasons." *Political Analysis* pp. 1–20.

Gaynor, SoRelle W, Kristina Miler, Pranav Goel, Alexander M Hoyle and Philip Resnik. 2022. "Do You Walk the Walk, Talk the Talk, or Tweet the Tweet? Legislators' Ideal Points Across Venues.".

Grand, Gabriel, Idan Asher Blank, Francisco Pereira and Evelina Fedorenko. 2022. "Semantic projection recovers rich human knowledge of multiple object features from word embeddings." *Nature human behaviour* 6(7):975–987.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21(3):267–297.

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.".

Herrnson, Paul S, Costas Panagopoulos and Kendall L Bailey. 2019. *Congressional elections: Campaigning at home and in Washington.* Cq Press.

Hirano, Shigeo, Gabriel S Lenz, Maksim Pinkovskiy and James M Snyder Jr. 2015. "Voter learning in state primary elections." *American Journal of Political Science* 59(1):91–108.

Jewitt, Caitlin E. and Sarah A. Treul. 2014. "Competitive primaries and party division in congressional elections." *Electoral Studies* 35:140–149.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American political science review* 97(2):311–331.

Lee, Frances E. 2009. *Beyond Ideology: Politics, Principles, and Partisanship in the U.S. Senate.* Chicago, IL: University of Chicago Press.

Mikolov, Tomáš, Wen-tau Yih and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies.* pp. 746–751.

Montagnes, Brendan Pablo and Jon C Rogowski. 2015. "Testing core predictions of spatial models: Platform moderation and challenger success." *Political Science Research and Methods* 3(3):619–640.

Nyhan, Brendan and Jacob M Montgomery. 2015. "Connecting the candidates: Consultant networks and the diffusion of campaign strategy in American congressional elections." *American Journal of Political Science* 59(2):292–308.

Poole, Keith T. and Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29(2):357–384.

Porter, Rachel, Maura McDonald and Sarah Treul. 2021. "Changing the Dialogue: Descriptive Candidacies and Position Taking in Campaigns for the US House of Representatives.".

Porter, Rachel and Sarah A. Treul. 2023. "Evaluating (In)Experience in Congressional Elections." `https://rachelporter.org/files/inexperience.pdf`.

Porter, Rachel, Sarah A. Treul and Colin R. Case. 2023. "Database on Primary Election Website Content." *Chapel Hill, NC; University of North Carolina Libraries* .

Ramey, Adam. 2016. "Vox populi, vox dei? Crowdsourced ideal point estimation." *The Journal of Politics* 78(1):281–295.

Rheault, Ludovic and Christopher Cochrane. 2020. "Word embeddings for the analysis of ideological placement in parliamentary corpora." *Political Analysis* 28(1):112–133.

Roberts, Jason M. 2007. "The Statistical Analysis of Roll-Call Data: A Cautionary Tale." *Legislative Studies Quarterly* 32(3):341–360.

Rodriguez, Pedro L and Arthur Spirling. 2022. "Word embeddings: What works, what doesn't, and how to tell the difference for applied research." *The Journal of Politics* 84(1):101–115.

Rodriguez, Pedro L, Arthur Spirling and Brandon M Stewart. 2021. "Embedding Regression: Models for Context-Specific Description and Inference." *American Political Science Review* pp. 1–20.

Shor, Boris and Nolan McCarty. 2011. "The ideological mapping of American legislatures." *American Political Science Review* 105(3):530–551.

Slapin, Jonathan B and Sven-Oliver Proksch. 2008. “A scaling model for estimating time-series party positions from texts.” *American Journal of Political Science* 52(3):705–722.

Sulkin, Tracy. 2005. *Issue Politics in Congress.* Cambridge University Press.

Tausanovitch, Chris and Christopher Warshaw. 2017. “Estimating candidates’ political orientation in a polarized congress.” *Political Analysis* 25(2):167–187.

Thomsen, Daniel M. 2022. “Competition in Congressional Elections: Money Versus Votes.” Presented at the 2020 Annual Meeting of the American Political Science Association.

Thomsen, Danielle M. 2014. “Ideological moderates won’t run: How party fit matters for partisan polarization in Congress.” *The Journal of Politics* 76(3):786–797.

Treul, Sarah A and Eric R Hansen. 2023. “Primary Barriers to Working Class Representation.” *Political Research Quarterly* p. 10659129231154914.

Vafa, Keyon, Suresh Naidu and David M Blei. 2020. “Text-based ideal points.” *arXiv preprint arXiv:2005.04232* .

Xing, Eric P and Tony Jebara. 2014. “Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32.”.