Estimating Pro-War and Anti-War Sentiments by State Using Multilevel Regression with Poststratification

Parker Kaukl

5/12/2022

Abstract

One of the most complex topics in American Politics is Military Intervention. Public Opinion on Military Intervention is historically divided into Pro-War and Anti-War positions. Understanding which states may be more or less likely to be Pro or Anti War is important for predicting which areas may have more protests, and inform policy positions for elected officials. This paper will create a model to predict opinions on war using a Frequentist and a Bayesian approach. Then, these models will be used in order to create a Multilevel Regression with Post-Stratification to estimate the public opinion within each state in the United States of America. In both the Bayesian and the Frequentist approaches, our Multilevel Regression with Post-Stratification estimates that the public opinion within each state to believe that War is effective is between 60% and 70%. These results tend to be similar to the public opinion polls on war, before any casualties have been incurred.

Background and Significance

In a democracy, politicians are accountable to their constituents. Every two years, all 435 Representatives in the U.S. House are up for reelection. Mansbridge (2003) describes that promissory responsiveness, whereby citizens respond to campaign promises and hold politicians accountable, is virtually impossible with ambiguous communication (Mansbridge, 2003). When candidates do not engage in explicit policy debate, they deprive voters of critical information which may keep them from fully participating in the democratic process or lead them to base their political choices on other, suboptimal criteria (Druckman et al. 2010). Unfortunately, early research by Downs showed that candidates and parties have incentives to becloud their policies in a fog of ambiguity (Downs, 1957). Scholars differ in their analyses of these incentives. Tomz and Houweling argue that under the right circumstances, voters actually "embrace" ambiguity because they make optimistic inferences about candidates' issue positions when campaigns send murky signals (Tomz and Houweling, 2000). In contrast, Krupnikov and Ryan note that preferences for ambiguous candidates are not equivalent to voting for those candidates (Krupnikov and Ryan, 2017). This is important to consider, as candidates' primary incentive is election; it doesn't matter if non-voters love you. Milita, Simas, Ryan, and Krupnikov (2017) add nuance to these findings, arguing that in the absence of issue statements, voters will rely on party heuristics to make inferences.

Measuring ambiguity is difficult, and explains why much scholarship on the topic has either avoided empirical testing altogether, or relied on small samples of communication on a limited range of issues (Chap et al. 2019). In this paper, we investigate the ambiguity of political rhetoric in 2014 United States congressional candidates using data from Chap et al. (2019). Their data allow us to analyze both the personal characteristics of candidates and the characteristics of their district, and how they relate to ambiguity. Following Milita, Simas, Ryan, and Krupnikov (2017), we test whether ambiguity differs with candidate party identification (Democrat or Republican) interacting with ideology. Candidates from one party may feel more comfortable with the idea that voters rely on party heuristics when faced with ambiguous language. Similarly, one party may fear that effect and seek to counteract it with more clear language. Following Shepsle, we investigate whether incumbents tend to have greater clarity than challengers, potentially because they are constrained by past voting behavior, greater media scrutiny, and previous position statements (Shepsle, 1972). We expect our model to agree with Chap et al. that as a district becomes more demographically heterogeneous, ambiguity will increase. This may be because more diverse districts will tend to have less ideological consistency, and therefore incentivize candidates to broaden their appeal. Finally, we consider whether more issue pages on a candidate's website are associated with clearer language. Candidates seek to prime their strengths and avoid their weaknesses, there is little incentive to address an issue ambiguously when avoidance will do (Drucker et al. 2004).

Data

The dataset used is from Chap et al. (2019) , which explored 2014 congressional candidates' ambiguity on political issues. The authors "hand-coded a random sample of 2012 congressional candidates' websites, assigning an ambiguity score." A total of 870 websites from 2014 were then automatically scored using Wordscores, a program designed for political textual analysis.

Wordscores is a textual analysis software that performs spatial regressions and frequency clouds on training texts to analyze the "ambiguity" of political statements by analyzing their distance from actionable positions as compared to adjectival statements about the issues. In the original paper that documented the creation of Wordscores, researchers stated that "because it treats words unequivocally as data, our technique not only allows us to estimate policy positions from political texts written in any language but, uniquely among the methods currently available, it allows us to calculate confidence intervals around these point estimates."

This kind of machine-learning analysis can be hard to interpret because verbs were shown to have much higher clarity scores in policy position pages, while adjectives and pronouns were seen as ambiguous. However, the verbs that were labeled as more clear were not necessarily actionable, such as "do," "can," and modifiers like "will." Relying on training texts (taken from the same genre of material) to code which words are more

or less ambiguous can be an unreliable measure, as other ambiguous political communication could just as easily be seen in the training texts. [Footnote 1]

The dataset is multilevel because it captures information both on a candidate and their district. The first level variables pertain to candidates. The candidates name, their party (a binary variable of 1 for Democrats and 0 for Republicans), incumbency (also binary, with 1 indicating incumbency), ideology and ambiguity are included. Ideology is a more specific measure of the candidate's left-right orientation. Higher (positive) scores indicate more conservative candidates and lower (negative) scores indicate more liberal candidates; this type of ideological measure was done on both candidates and their district to create a 'mismatch' variable. Mismatch is the distance between the candidate's ideology and the district's ideology. The response variable is ambiguity. It is an assigned score, produced by the Wordscores software; higher scores indicate greater ambiguity (i.e., less clarity). This score is a standardized measurement of ambiguity, where a measurement of 0 represents that a candidate's ambiguity is equal to the average ambiguity for all candidates. [Footnote 2]

The second level is the congressional district that the candidate is in. Each district is assigned an identification number. The district's ideological lean was measured to indicate whether the district was more liberal or conservative. There are also two "heterogeneity" variables at the district level, which measure the variability within a district's demographic information and ideological lean of their citizens. For both, higher scores imply more heterogeneity among voters.

To clean our data, we reversed the sign of ambiguity. In the original dataset, higher ambiguity scores indicated less ambiguity. This method of measurement was confusing, and therefore we decided to multiply the ambiguity value by negative one, to get a new variable. Because this variable was a z-score, this will only change the direction, but not the magnitude. In addition, we also removed a datapoint. In District 409, some of the values for district level measurements had values that were mathematically impossible, such as having a negative value of Attitude Heterogeneity. A value of 0 indicates that there is complete Attitude Homogeneity, so therefore, a negative value would be impossible.

Table 1: First 6 rows of the dataset

	Name	District ID	Ideology Score	Issue Pages	Democrat	District Mis-match	Incumbent	Demographic Hetero-geneity	Attitude Hetero-geneity	District Lean	Ambiguity
3	LEFLORE, BURTON R	101	0.3514	1	1	0.1818866	0	0.5845018	1.272789	0.3672764	-0.0156
4	BYRNE, BRADLEY ROBERTS	101	1.0819	7	0	0.5486134	1	0.5845018	1.272789	0.3672764	-0.3933
5	WRIGHT, ERICK	102	-0.0038	6	1	0.6848035	0	0.5906023	1.188103	0.5017814	-0.3800
6	ROBY, MARTHA	102	0.7752	0	0	0.0941965	1	0.5906023	1.188103	0.5017814	0.4444
7	SMITH, JESSE TREMAIN	103	0.0104	4	1	0.6099995	0	0.5841383	1.224224	0.4465979	-0.4260
8	ROGERS, MICHAEL	103	-1.1222	0	0	1.7425995	1	0.5841383	1.224224	0.4465979	-0.2193

We start with a correlation plot (Figure 1) and a scatterplot matrix (Figure 2) to analyze the relationships between numeric and continuous variables. As shown in Figure 1, the strongest linear correlation is between ideology and partisanship, which is expected because both are measures of ideology in their own way and (with few exceptions), democratic candidates will fall on the left and Republican candidates on the right.
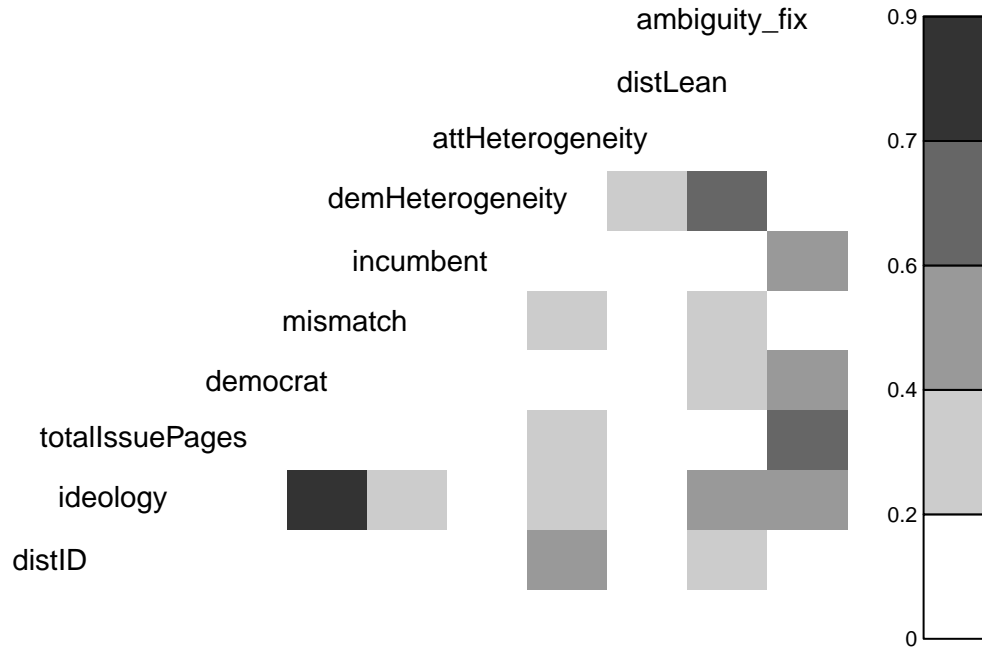
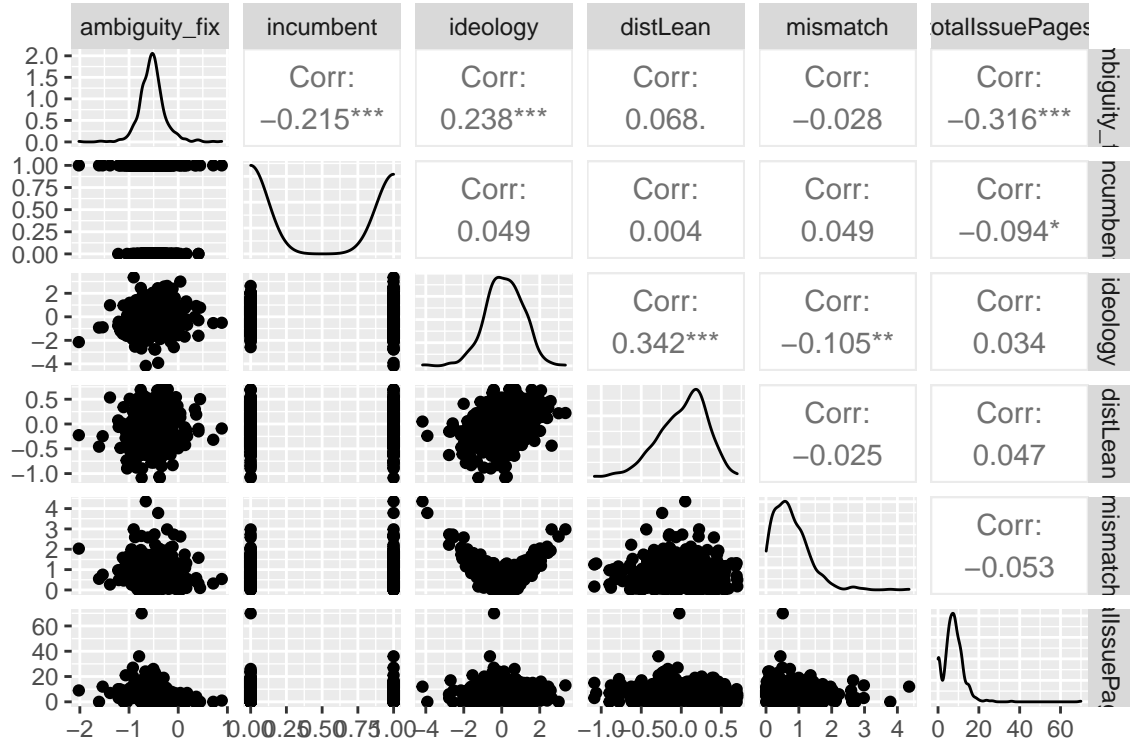Figure 1: Correlation Plot of Numeric Variables in the Dataset

Figure 2: Correlation Matrix of Continuous Variables in the Dataset

4

Following this, we specifically explore the relationships between ideological variables `ideology`, `party`, and `incumbency`. Figure 3 shows an overall trend of more ambiguous political communications as a candidate becomes more ideologically conservative. Figure 2 shows a more nuanced picture of the relationship between `ideology` and `ambiguity` where slopes differ according to party.The tendency for both parties is to have more ambiguous language for more candidates with moderate ideologies, but the effect is more extreme for Democrats. When we add whether a candidate was an incumbent or challenger in Figure 4, we see that for Republicans `incumbency` has a balancing effect. We see consistent ambiguity scores for Republican incumbents over a range of ideologies, unlike the other three groups. From Figure 5 we can also see that for both parties, challengers have more ambiguous language overall than their incumbent counterparts. This is reflected in Table 1 comparing `mean_ambiguity` for each group. These figures suggest we should consider adding interaction terms for `party` and `ideology`, and `party` and `incumbency`. Figure 6 shows the distribution of ideologies by party. We see overlap near the center of the distribution. This histogram, along with the party/ideology correlation test result showing correlation = -0.6724306 tell us that we should proceed with models including party and evaluate further using AIC and BIC.
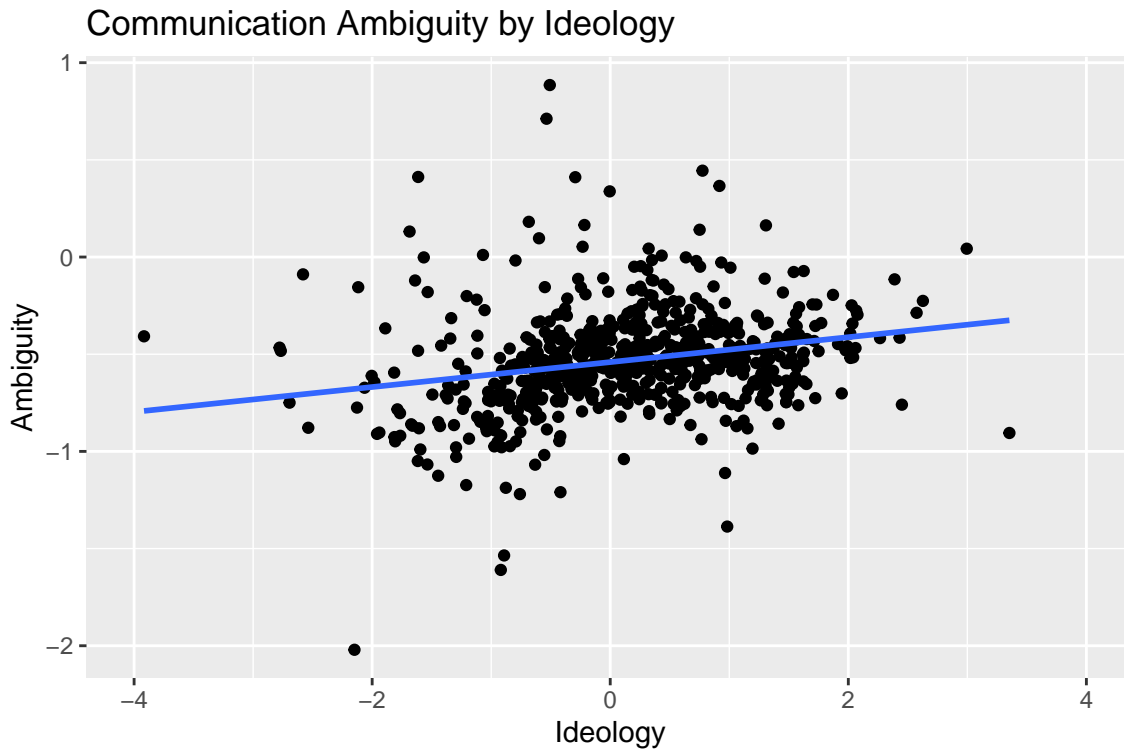

Figure 3: Scatterplot of Communication Ambiguity by Ideology

This exploratory analysis of the relationship between party, ideology, and their effect on ambiguity was particularly important due to the strong but incomplete correlation of partisanship and ideology. We felt that using both ideology and party as a fixed effect may cause multicollinearity when we were originally creating our model. However, our correlation tests, plots, and further analysis have suggested that using both variables would not create a multicollinearity problem. In addition, both variables play an important part in profiling a candidate and explaining ambiguity in their messaging. Table 2 shows that incumbent Republicans and Democratic challengers have very similar mean ambiguity scores.
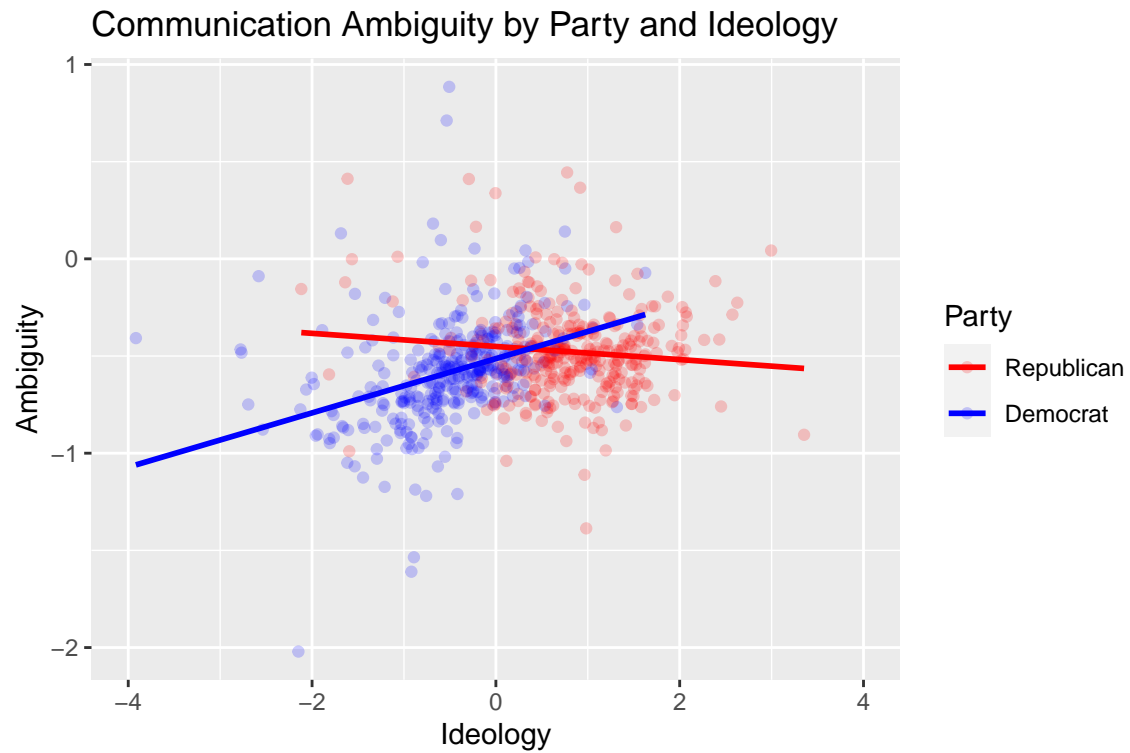
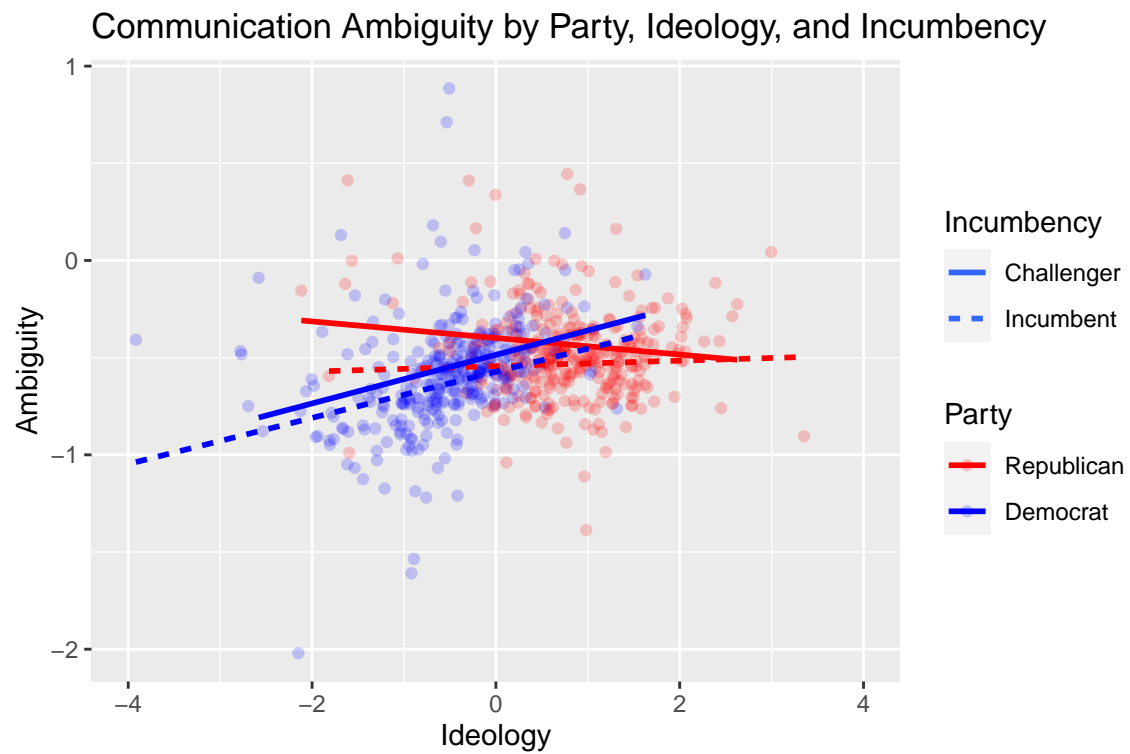Figure 4: Scatterplot of Communication Ambiguity by Ideology, Grouped by Partisanship

Figure 5: Scatterplot of Communication Ambiguity by Ideology, Grouped by Partisanship and Incumbency
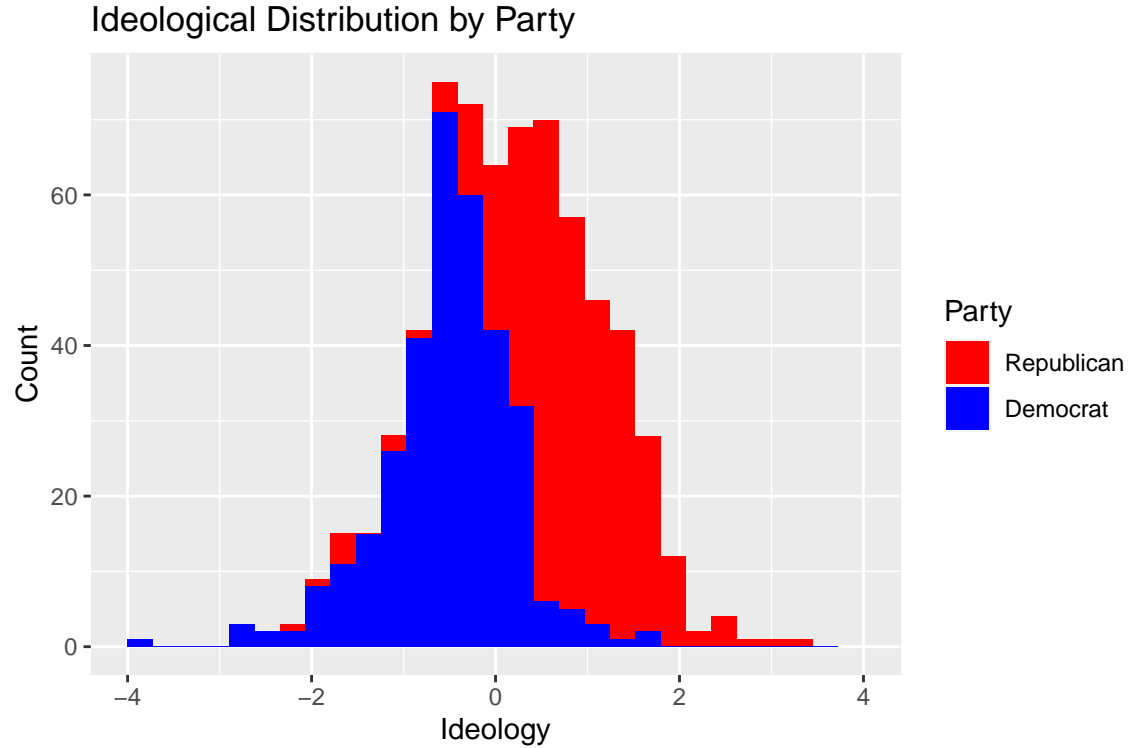
Ideological Distribution by Party

Figure 6: Ideological Distribution by Party

Table 2: Summary Statistics of Ambiguity, Grouped by Partisanship and Incumbency

Democrat	Incumbent	Mean Ambiguity	Standard Deviation of Ambiguity	Minimum Ambiguity	Maximum Ambiguity	Number
0	0	-0.4231988	0.2134283	-1.0396	0.4120	161
0	1	-0.5294059	0.2415471	-1.3868	0.4444	170
1	0	-0.5255372	0.1900786	-1.2198	0.0434	188
1	1	-0.6633632	0.3457776	-2.0206	0.8850	144

One important dimension of the political data set focuses on the political characteristics of the district. Some of these variables are important to predict political ambiguity. For instance, we might predict that districts that are more polarized might have more ambiguity in their politicians. In districts with higher levels of polarization, or in districts that have a greater variety in political ideologies, we might expect that there is a different level of ambiguity in a politician. Therefore, it may be important to observe different district level variables.

First, we will construct numerous plots to observe the different variables at a district leve, in order to understand their distribution and the relationship that these variables have with ambiguity. These resulting figures are shown in Figure 7. District Lean, Attitude Heterogeneity and Demographic Heterogeneity all have a roughly normal distribution. District Demographic Heterogeneity and District Lean both appear to have skews, with District Lean having a skew towards liberal ideologies, and a skew towards districts with more Demographic Heterogeneity. In these graphs, we are able to see a few weaker, general trends. We can see that there is no clear correlation between ambiguity and District Lean, or District Attitude Heterogeneity and Ambiguity. However, the District Demographic Heterogeneity appears to have a relationship with the ambiguity, so we might consider adding that variable.
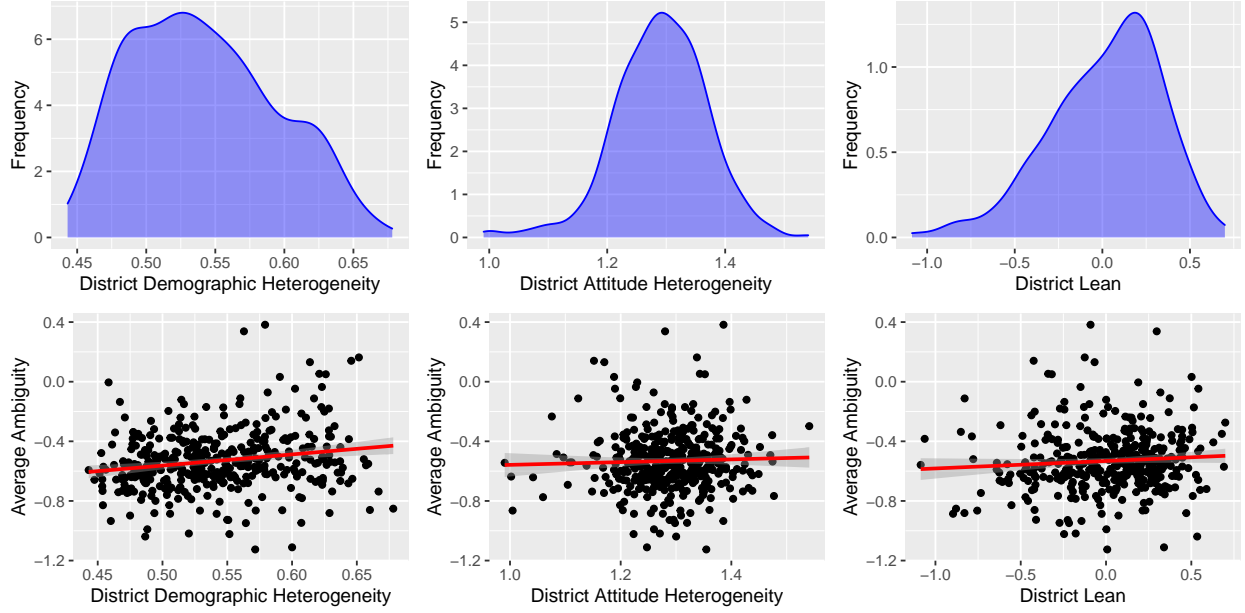
Figure 7: Analyzing Frequency and Trends of District-Level Variables in Relation to Ambiguity

Methods

To evaluate different models, we have used the AIC and BIC values. These values show us how much variability in a model is accounted for, while penalizing the use of additional variables. We decided that the best model minimizes these AIC and BIC terms compared to other models. In addition, the model that we want to choose would investigate the relationships with ambiguity that we have previously identified. We started with a baseline unconditional means model. From there, we decided to include additional terms, as this improved the AIC and BIC of our model. An ideal model for this report would reduce the AIC and BIC relative to other models, while addressing the effect on ambiguity of candidate's party, ideology, the number of issue pages and demographic heterogeneity.

With only two observations per district, any random slope would be created from only those two observations and thus any random slope term would be inappropriate to apply to this dataset. Therefore, we choose to use a random effect term for District, instead of any other random effects or slopes.

First, we identified that we wanted to use a model that included ideology and the incumbent status of the candidate. Then, we additionally recognized that the variable for demographic heterogeneity was important in predicting the ambiguity of a candidate's position. Figure 7 shows a positive correlation between District Demographic Heterogeneity and Ambiguity. We also found a relationship between the number of issue pages and the ambiguity of a candidate. This model minimized AIC and BIC more than any other model that we tried.

Finally, we included an interaction term between ideology and the candidate's party. This was based on Figure 4, which shows a relationship between the candidate's party and ideology. Therefore, we have a model, where i represents district and j represents the candidate:

This will create a model that has an AIC of approximately -42.68, and a BIC of approximately -2.209. These are remarkably small values for AIC and BIC compared to the other models that we tried to make. Some models had slight improvements on AIC, but these results were often only a small improvement on AIC, and a significant reduction in effectiveness for BIC. In addition, this model tested the relationships that we wanted to investigate.

$$Y_{ij} = [\alpha_0 + \beta_0 \text{ideology}_{ij} + \beta_1 \text{ideology}_{ij} \times \text{democrat}_{ij} + \beta_2 \text{democrat}_{ij} + \beta_3 \text{incumbent}_{ij} \tag{1}$$

$$+ \beta_4 \text{Dem.Heterogeneity}_i + \beta 5 \text{TotalIssuePages}_{ij}] + [u_i + \epsilon_{ij}] \tag{2}$$

where, $u_i \sim N(0, \sigma_u^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$.

α_0 is the variable that indicates the ambiguity we expect for challengers who are republican, in districts with a value of demographic heterogeneity of zero and there are no pages. This is not a valuable measure, because we do not expect there to be a case where this reflects a real scenario.

β_0 represents the expected change that we would see in ambiguity scores for a one-point increase in ideology score for republicans, all other variables held constant. Likewise, β_1 represents the expected change that we would see in ambiguity scores for a one-point increase in ideology score for democrats, all other variables held constant. β_2 represents the difference between democrats and republicans for candidates with ideology scores of zero, all else held equal.

β_3 represents the expected change that we would see in ambiguity scores for incumbents, when compared to challengers, all else held equal. β_4 represents the change that we would expect in ambiguity scores as the Demographic Heterogeneity increases by a score of one within a district. Finally, β_5 represents the change in ambiguity scores as the total number of issue pages is increased by one for a candidate, all else held equal.

Results

Table 3: Variance and Covariance Estimates for Random Effects

Group	Var. 1	Var. 2	Variance	Standard Deviation
distID	(Intercept)	NA	0.0001	0.0079
Residual	NA	NA	0.0504	0.2244

Table 4: Estimates of Fixed Effects

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-0.7493	0.1001	435.5119	-7.4831	0.0000
ideology	0.0129	0.0174	654.7179	0.7434	0.4575
democrat1	-0.0479	0.0238	555.7621	-2.0118	0.0447
incumbent	-0.1175	0.0186	495.6615	-6.3230	0.0000
demHeterogeneity	0.8153	0.1768	409.8686	4.6115	0.0000
totalIssuePages	-0.0155	0.0017	655.7906	-9.0158	0.0000
ideology:democrat1	0.1010	0.0253	648.0790	3.9880	0.0001

The model estimates that Republican challengers with no issue pages, an ideology of 0, running in a district with no demographic heterogeneity will have communication ambiguity of -0.749305. For each additional point in ideology for Republicans, the model estimates a 0.012945 point increase in communication ambiguity after accounting for all other fixed effects. However, it should also be noted that this result is not statistically significant, with a large p-value of 0.45. Democrats are estimated to have communication ambiguity 0.047871 less than Republicans after accounting for all other fixed effects. Incumbents are estimated to have communication ambiguity 0.117505 less than challengers after accounting for all other fixed effects. For each additional point in a district's demographic heterogeneity, the model estimates a 0.815260 point increase in communication ambiguity after accounting for all other fixed effects. For each additional issue page on a candidate's website, the model estimates a 0.015542 point decrease in communication ambiguity

after accounting for all other fixed effects. For each additional point in ideology, the model estimates that Democrats have an additional 0.100959 point increase in communication ambiguity over Republicans. The model estimates that the standard deviation in communication ambiguity across districts is 0.007862 after accounting for all fixed effects.

Discussion and Conclusions

In this paper, we saw that ambiguity does differ with an interaction between a candidate's party identification and their ideology. Given a p-value of roughly zero for `ideology:democrat`, we feel confident that Republicans have a different change in ambiguity depending on ideology than Democrats. The estimate for this variable is 0.100959 and indicates that Democrats become more ambiguous per increase in ideology than Republicans. This connects to the theory from Milita, Simas, Ryan, and Krupnikov (2017) that depending on party, candidates may feel more comfortable with the idea that voters rely on party heuristics when faced with ambiguous language. Our results add evidence to Shesle's research that incumbents tend to have greater clarity than challengers overall, potentially because they are constrained by past voting behavior, greater media scrutiny, and previous position statements (Shepsle, 1972). The model estimate for `incumbent` is -0.117505, with a p-value of nearly zero, indicating we can be very confident incumbents tend to be more clear. Our model does agree with and provide further evidence to Chap et al. that as a district becomes more demographically heterogeneous, ambiguity increases. We see an estimate of 0.815260 for `demHeterogeneity` and a p-value that was roughly zero. Finally, we see statistical evidence of a relationship between the number of issue pages on a candidate's website and less ambiguous language, albeit not a dramatic decrease. The estimate for `totalIssuePages` is -0.015542 with a p-value of approximately zero. This answers our last research question following (Drucker et al. 2004) who concluded that there is little political incentive to address an issue ambiguously when avoidance will do. When candidates do address an issue, they are more likely to be clear about it.

Studies of ambiguity in political communication could benefit from a greater understanding of political success with our research questions in mind. Since the primary driver for most politicians is to win the election, it would be fascinating to consider vote shares between candidates. Other interesting directions for this research would be to compile data from many more election cycles to track individual candidates, parties, and other variables over time. This could add nuance to interactions between party and ideology depending on broader public opinion on generic ballots and control of the white house. For example, a party in 2014 may have been okay with voters relying on heuristics, if the public attitudes towards them were favorable, and opted for more ambiguous language. But generally opinions change for parties and candidates, especially depending on who controls the white house.

It is clear that certain candidate- and district-level factors lead to more ambiguous political communications. Going forward, it may be beneficial to our democracy to consider what, if any, constituent-induced factors contribute to the ambiguity of a candidate's communications. That research may give greater agency to voters to shape political discussions and help strengthen our democracy.

References

Chapp, C., Roback, P., Johnson-Tesch, K., Rossing, A., & Werner, J. (2019). Going Vague: Ambiguity and Avoidance in Online Political Messaging. Social Science Computer Review, 37(5), 591–610.

Downs, A. (1957). An economic theory of democracy. New York: Harper.

Druckman, J. N., Hennessy, C. L., Kifer, M. J., & Parkin, M. (2010). Issue engagement on congressional candidate web sites, 2002—2006. Social Science Computer Review, 28(1), 3-23.

Druckman, J. N., Jacobs, L. R., Ostermeier, E. (2004). Candidate strategies to prime issues and image. The Journal of Politics, 66, 1205–1227.

Krupnikov, Y., Ryan, J. B. (2017). Choice vs. action: Candidate ambiguity and voter decision making. Quarterly Journal of Political Science, 12, 479–505.

Mansbridge, J. (2003). Rethinking representation. American political science review, 97(4), 515-528.

Milita, K., Simas, E., Ryan, J., Krupnikov, Y. (2017). The effects of ambiguous rhetoric in congressional elections. Electoral Studies, 46, 48–63.

Shepsle, K. A. (1972). The strategy of ambiguity: Uncertainty and electoral competition. The American Political Science Review, 66, 555–568.

Tomz, M., Van Houweling, R. P. (2009). The electoral implications of candidate ambiguity. American Political Science Review, 103, 83–98.

Appendix

[1] For more information on the specific methodology of their textual analysis, the original paper of Word-scores may be of interest. From that paper: "The ultimate methodological price to be paid for the benefits of a posteriori interpretation is the lack of any objective criterion for deciding between rival spatial inter-pretations, in situations in which the precise choice of interpretation can be critical to the purpose at hand. The price for taking the a priori route, on the other hand, is the need to accept take-it-or-leave-it proposi-tions about the number and substantive meaning of the policy dimensions under investigation. Using the a priori method we introduce here, however, this price can be drastically reduced. This is because, once texts have been processed, it is very easy to re-estimate their positions on a new a priori dimension in which the analyst might be interested. For this reason we concentrate here on estimating positions on a priori policy dimensions."

[2] In some of our data wrangling, we chose to switch the sign of the ambiguity variable in the dataset. Previously, it was higher scores that indicated higher clarity. However, because ambiguity is the response variable and the variable is titled 'ambiguity' we thought it made more sense for analysis to switch these around.

This is important for two reasons: First, any outside analysis that a person sees about Wordscores software will seem to draw the opposite conclusions; this is not the case, but the language prioritizes clarity over ambiguity. And second, this mutation was made after beginning work on our graphics and may not be reflected as clearly as possible in our analysis. This footnote serves to hopefully dispel any confusion.