An Experimental Approach to Measuring Ideological Positions in Political Text*

John A. Henderson†

Assistant Professor

Political Science

Yale University

August 30, 2016

Abstract

Though powerful as general tools, automated measures of position-taking in text often perform poorly when models of speech are difficult to develop or theoretically contested. Rather than model text, I develop an experimental approach to measure perceptions of partisanship in speech, with an application to 2008 Congressional advertisements. I randomly assign ads to subjects recruited in a large-N survey, and ask them to 'guess' the party of featured candidates, with ads scored as their average party inference. These *party perception scores* are empirically synonymous with a liberal-conservative dimension, and highly reliable across samples and experimental conditions. Party identity has little impact on guesses, indicating the inferential task significantly mutes partisan bias. For validation, I assess which words influence guessing, and whether ad-scores correspond to expectations about how candidates target voters. Importantly, this experimental approach can augment or validate automated text analysis, and generalize to study speech across many other contexts.

---

*For valuable comments I thank John Bullock, Devin Caughey, Stephen Goggin, Jacob Hacker, Greg Huber, Stephen Jessee, Michael Laver, Burt Monroe, Ellie Powell, Arthur Spirling, Alex Tahk, Alex Theodoridis, Jonathan Wand, Chris Wlezian, and also participants in workshops at UT-Austin, Wisconsin and Yale. All errors are my responsibility.

†<john.henderson@yale.edu>, http://www.jahenderson.com, Institution for Social and Policy Studies, Yale University, 77 Prospect Street, New Haven, CT 06520

1 Introduction

Over the last two decades, there has been dramatic growth in the use of text data to measure the ideological leanings of parties, candidates and voters (Benoit et al. 2009; Grimmer and Stewart 2013; Slapin and Proksch 2008). Though early pioneers utilized coders to analyze the content of text, recent work has turned to powerful automated approaches to scale word usage on a common space (Lauderdale and Clark 2014; Laver et al. 2003; Slapin and Proksch 2008). Yet, scaling speech automatically has proven to be challenging in many contexts (e.g., Lauderdale and Herzog 2016; Quinn et al. 2010). Current models of political behavior, including utility accounts of legislative voting, often transport imperfectly to the domain of word choice, especially when speech is insincere or institutionally unconstrained.[1] Compounding this difficulty, scholars frequently disagree about which theoretical accounts explain how politicians communicate to the public, and thus what metrics can serve as appropriate benchmarks for scaling speech. Consequently, in evaluating a new approach, disputes over whether a novel measure is valid can become indistinguishable with whether a particular theory of political communication is correct.

Instead of modeling the latent structure of words, I develop an experimental approach that taps collective perceptions to directly measure the content of political speech. I randomly assign political statements to subjects, in a large-N survey, and ask them to infer particular features from the messages. In this study, I present survey subjects with statements drawn from U.S. House and Senate campaign advertisements, and ask them to *guess* the party (or ideological leaning) of the candidate featured in each ad. Ads are then scaled as their average *party perception score*. Unlike previous survey-based approaches that exhibit bias due to how co-partisans locate politicians or messages in ideological space (Ansolabehere and Brady 1989), I find the inferential nature of the party guessing task virtually eliminates partisan biases in measuring ads. By placing people behind

[1]Sparsity can also make it difficult to model complex uses of language, including word interactions, e.g., "woman's right to choose" vs. "woman's right to choose her doctor."

a veil of uncertainty, the design taps the independent wisdom of crowds. Consequently, this scaling approach is remarkably reliable in both representative and non-representative samples, and across a variety of experimental conditions. Further, replacing party with ideological labels, I show that party perception scores are indistinguishable from a liberal-conservative dimension in the mind of the American voter.

I implement this party guessing design in three waves. For the first, I recruit a representative sample of 1,800 respondents in the 2014 Cooperative Congressional Election Study (CCES) to scale 150 positive and negative Congressional ads sampled from the 2008 election cycle. In a second wave, I replicate the party guessing experiment using a non-representative sample totaling 4,853 subjects recruited through Amazon Mechanical Turk (MTurk). In this subject pool, I scale 50 ads from the above 150, and an additional 50 ads left out of the original CCES study (totaling 200 ads), to evaluate how well the guessing task replicates across dramatically different conditions. Beyond assessing whether guessing depends on sample demographics (e.g., political interest, party identity), I also conduct additional experiments in this MTurk wave. These include randomizing positive or negative ads to be guessed first, varying the expected number of guesses per ad, and offering monetary payments and penalties for correct or incorrect responses, among others. The resulting scores are very similar across all conditions.

On the basis of these results, in a final guessing wave, I recruit a highly unrepresentative MTurk sample to score a much larger set of 2,103 general election ads aired in 2008 as collected by the Wisconsin Ads Project (CMAG). Subjects were allowed to evaluate as many ads as they wanted, and though statements were randomly ordered across survey tasks, these were not randomly assigned to subjects. Nevertheless, perception scores are again virtually identical for the 200 ads that overlap in the above waves, strongly confirming the replicability of this inferential mode of scaling campaign advertising.

In the second half of this study, I develop a number of tests to validate party perception scores against two automated scaling alternatives: *Wordfish* and *Wordscores*

(Laver et al. 2003; Slapin and Proksch 2008). I investigate which words are most predictive within each scaling approach to assess whether these scorings are sensible. I then conduct vignette experiments in the CCES, presenting real candidates airing one of 75 randomly selected positive ads. In the vignettes, subjects are asked to place candidates on a liberal-conservative scale, given the ad statements and some additional policy information. Notable to the design, the randomized ads were previously scored through party guessing by *an entirely different sample of survey respondents.* The results show subjects updated their ideological perceptions of the candidates in line with ad party perception scores, but not as predicted by *Wordfish* or *Wordscores* scalings, indicating party guessing is a better measure of the policy information voters actually access and incorporate.

One major contribution of this study is the construction of new measures of candidate position-taking in a large number of ads from the 2008 election. Using this larger sample of ads measured in wave three, I illustrate how party perception scores can be used to explore a number of central theories about electoral advertising, including whether campaigns converge over the cycle to target median voters or consistently emphasize their partisan or positional differences. For this, I measure the correlation between ad party guesses, district presidential vote, and scaled roll calls. As an additional validation, I compare this to analogous correlations using *Wordfish* and *Wordscores*, finding that party perception scores are consistently correlated with district presidential vote and roll call positions, while the automated estimates are inconsistent predictors of both. Lastly, I explore a number of ways in which the party guessing approach can be incorporated into machine learning and the automated scaling of text data, and propose extensions to measuring ideology going beyond political advertising in the American two-party context.

2 Prior Approaches to Scaling Speech

Scaling the behaviors of political elites on an ideological dimension has had a long pedigree in political science (Clinton et al. 2004; Poole and Rosenthal 1997). Early on,

researchers utilized survey respondents to locate political actors on some pre-defined ideological scale (e.g., 'Very Liberal' to 'Very Conservative'). Yet, among other concerns, motivated partisans would often evaluate in- and out-party politicians in expressive ways, biasing the resulting measures (Ansolabehere and Brady 1989). An important innovation was to move away from having biased or imperfectly-informed voters make judgements about politicians' positions, and to instead infer these using a utility model of legislative choice (Clinton et al. 2004; Poole and Rosenthal 1997). From the basic premise that legislators support proposals closer to their most preferred policy than the existing policies these aim to replace, researchers have produced a powerful measurement tool that can reliably summarize a great deal of the conflict within and across Congresses.

The analysis of position-taking in text data has taken a parallel course, with much of the early work employing coders to rate various dimensions in political documents (Benoit et al. 2009; Feinstein and Schickler 2008; Gerring 2001; Riker 1996). Yet, the recent explosion in the availability of political texts has led scholars to substitute labor-intensive content analysis with efforts to automate the granular analysis of text-data. This automated analysis of political text generally takes one of two forms. An *unsupervised* approach extends the utility framework of legislative voting into the domain of political speech, modeling words as being used based on how 'close' they are to describing a person's ideal policy position (Monroe and Maeda 2004; Slapin and Proksch 2008). Alternative *supervised* methods score words based on how well they predict an annotation (e.g., party), with documents scaled as their expected 'prediction score'.

A widely-used implementation of this unsupervised utility approach is the *Wordfish* model developed by Slapin and Proksch (2008). Accordingly, speech takes the form:

$$\begin{aligned} W_{ij} &\sim Poisson(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\rho_i + \psi_j + \beta_j \times \mu_i). \end{aligned} \tag{1}$$

The Poisson parameter, λ_{ij}, increases in the number of times legislator i speaks word

j, measured by W_{ij}. The term ρ_i measures the verbosity of i, while ψ_j measures the obscurity of j. The term β_j measures the amount of discrimination in j, or the degree to which the word is likely to be used mostly by liberals rather than conservatives. This word discrimination parameter plays an important role, tuning how much influence the legislator ideal point μ_i plays in determining the frequency that i speaks word j.

Supervised methods differ in an important way from unsupervised ones in aiming to score texts based on the empirical association between the string of words W_i used by i, and some document outcome P_i, typically party. One groundbreaking variant of this method, *Wordscores*, builds a dictionary of 'liberal' and 'conservative' words, and scores texts based on the frequency in which these ideological words are used (Laver et al. 2003). Reference texts D and R are first chosen by the researcher to represent canonical liberal and conservative statements. The target is to estimate $p(D|W_j)$ and $p(R|W_j)$, or the probability a document is liberal or conservative given the use of W_j in these pre-determined statements. Following Beauchamp (2010), denote $W_j^{(R)}$ to be the count of W_j appearing in document R, and analogously for $W_j^{(D)}$. Word proportions, capturing the frequency word j appears in a conservative text, are first constructed:

$$p(R|W_j) \;=\; \frac{W_j^{(R)}}{W_j^{(R)} + W_j^{(D)}}.$$

Then a word score B_j is built (assuming a word weight and polarity of ± 1) as $B_j = p(R|W_j) - p(D|W_j)$. Individual word scores are then used to scale the ith document as

$$\mu_i \;=\; \sum_j \frac{W_{ij}}{|W_i|} \times B_j, \tag{2}$$

where W_{ij} is the count of word j in i, and $|W_i|$ is the total number of words in i.

Since their development, both frameworks have been innovative and influential in political science (Grimmer and Stewart 2013; Lowe 2008; Monroe et al. 2008). Yet, each method has important limitations. Unsupervised utility models offer an awkward fit for

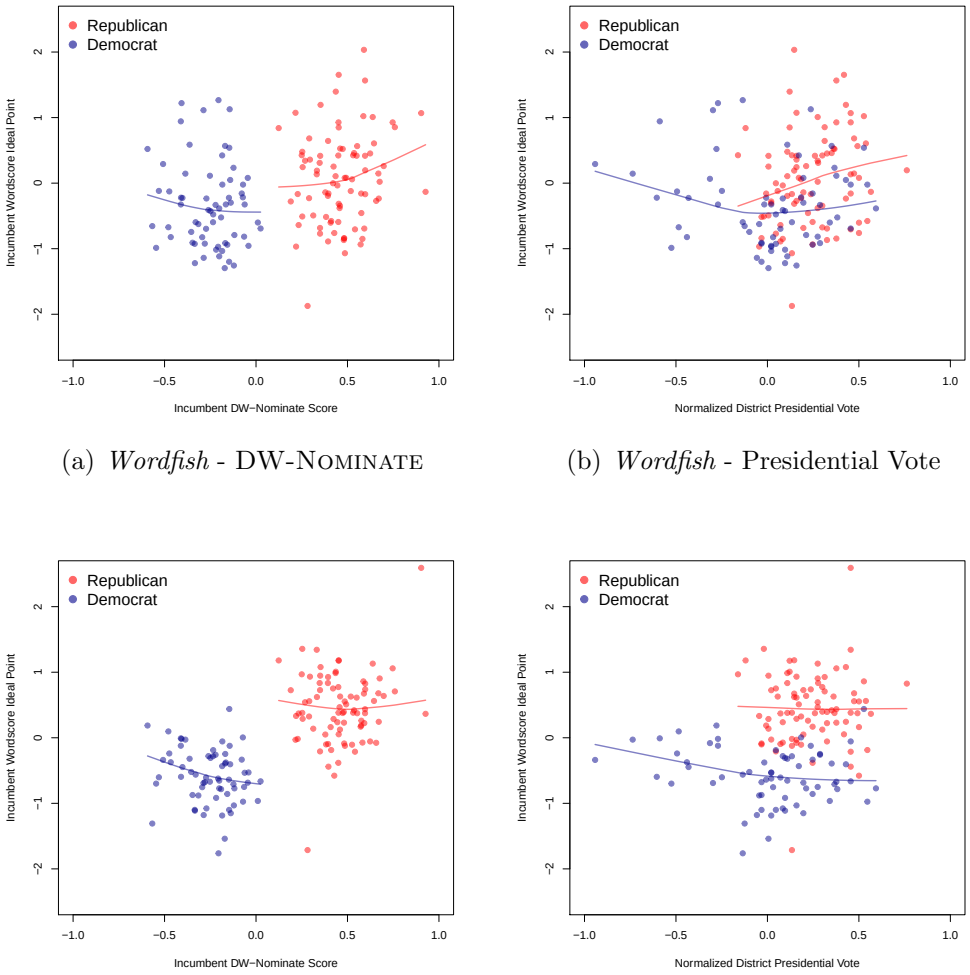
studying political speech. Further, these models will not work in all settings, including when speech is strategic or words convey little ideological meaning.[2] Alternatively, *Wordscores* assumes words have equal discrimination weight, and thus can overfit data (Lowe 2008; Monroe et al. 2008). It also eschews prior information to help smooth estimates when words exclusively appear in only conservative or liberal documents.[3] As practical heuristics, however, both approaches have proved quite successful in a number of contexts, especially scaling ideology in party platforms (Grimmer and Stewart 2013; Laver et al. 2003; Lo et al. 2014; Slapin and Proksch 2008).

2.1 The Validation Tautology

Though automated scaling methods differ in core assumptions, all fundamentally rely on some form of validation to assess the quality of their estimates (Benoit et al. 2016; Grimmer and Stewart 2013; Laver et al. 2011; Lowe and Benoit 2013). Further, this validation step *always* requires substantive judgements or strong theoretical assumptions that generally cannot be tested (e.g., Benoit and Laver 2007; Budge and Pennings 2007). A common validation strategy is to assume some prior ideological scale is *the* correct benchmark that accurately depicts political behavior in some domain, and to compare the novel measure against this baseline. For legislatures, the standard benchmark is DW-NOMINATE, or similar ideal point scalings of roll call voting (Poole and Rosenthal 1997). Scholars also frequently use measures of voter opinion (e.g., district presidential vote), to assess whether a new legislative scale meaningfully reflects the policy responsiveness thought to derive from the electoral connection (Tausanovitch and Warshaw 2013).

[2]If politicians avoid discussing policy then words will have no discrimination (Stokes 1992; Tomz and Van Houweling 2009). More broadly, it is unclear how well speech can be modeled generally in terms of preferences, which emerge from axioms of choice.

[3]Recently scholars have adapted these to be more fully Bayesian, including developing flexible models of political language (Lauderdale and Herzog 2016; Monroe et al. 2008).



(a) *Wordfish* - DW-NOMINATE (b) *Wordfish* - Presidential Vote

(c) *Wordscores* - DW-NOMINATE (d) *Wordscores* - Presidential Vote

Figure 1: **Correlation Between Automated Scalings of 2008 Congressional Advertising and Roll Call Positions or Presidential Party Vote**: Data are transcripts of 1,165 *positive* House and Senate commercials for the 2008 election collected by the Wisconsin Ads Project (CMAG). Ads are parsed into 1-gram words, stemmed and pruned of stop-words, then scaled using *Wordfish* and *Wordscores* models, with scores aggregated to the candidate-level. Roll call voting is scaled using standard DW-NOMINATE, and Presidential Vote is (normalized) state- or district-level proportion of party vote for president.

To illustrate this validation approach, I present such a comparison in Figure 1, using standard *Wordfish* and *Wordscores* scaling methods to estimate campaign positions from the text of congressional advertising. For data, I parse and scale the transcripts of 1,165 *positive* House and Senate ads aired by incumbents in the 2008 election.[4] Figures 1(a)

[4]Data come from 3,255 positive, contrast and negative ads aired in the top-210 media

and 1(c) plot *Wordfish* and *Wordscores* estimates against incumbent DW-NOMINATE roll call measures, while Figures 1(b) and 1(d) compare ad-scores to 2008 state or district presidential voting. Using these scores, we seen apparently little association between campaign advertising and either a legislator's record or her district's policy attitudes.

Ultimately, this validation is tautological. These comparisons necessarily assume that roll call records or voter attitudes are the appropriate benchmarks to assess the behavior of candidates in campaign mode. The fundamental challenge is that these weak correlations cannot tell us whether the text-based scales are poorly measured, or that it is wrong to assume campaign positioning is meant to reflect legislative behavior or voter opinion.[5] There are long-standing debates over whether campaigns avoid discussing issues (Stokes 1992), or appeal to their more extreme primary supporters (Snyder and Ting 2002), rather than faithfully discussing a candidate's record or targeting centrists (Henderson 2015). To verify the correct assumption among these alternatives, we require additional information that a particular scaling of speech is itself valid. But, this would obviate the need for further validation using benchmarks. In this study, I propose an altogether different approach, tapping independent human judgement to scale campaign speech.

---

markets in 2008 as collected by CMAG. Data described in more detail below.

[5]Similar concerns arise elsewhere. Lauderdale and Herzog (2016) and Quinn et al. (2010) find that certain unsupervised methods do a poor job at measuring ideological positions in legislative speech, given the amount of non-policy discussion that occurs on the floor. Both studies innovate by incorporating additional model structure to improve ideal point or topic measurement, though it is unclear similar structuring would be effective in the more free-wheeling campaign context. See the Appendix for scaling approach that directly incorporates additional theoretical structure in models of campaign speech.

3 Using Partisan Inferences to Scale Political Speech

The experimental framework I develop uses collective voter perceptions to scale political speech on an ideological dimension. A large number of respondents are shown randomly selected political texts, and asked to infer a politician's party from the statements. The design fits within a broader, burgeoning field of crowdsourcing to efficiently analyze data by distributing narrow rating tasks widely to a large number of coders (Benoit et al. 2016; Budak et al. 2016; Honaker et al. 2013; Lowe and Benoit 2013; Ororbia II et al. 2015). Scholars frequently crowdsource to validate automated analyses of text (Benoit et al. 2016; Lowe and Benoit 2013). Though increasingly, massive coding tasks are being used in political science to produce novel measures of media coverage (Budak et al. 2016; Ororbia II et al. 2015), campaign persuasion (Henderson 2015), democratization (Honaker et al. 2013), and sentiment in speech (Montgomery and Carlson 2016). In this work, crowdsourcing is most effective when average voters (rather than experts) are the intended recipients of speech, and when coding instruments are clear and concise (Ororbia II et al. 2015). Moreover, crowdsourcing offers particular advantages over machine learning when speech is sparse (e.g., in thirty-second ads), since people do not rely on the degree of co-occurrence between words to grasp meaning.[6]

In this application, I tap coder inferences to analyze position-taking in campaign advertisements. Formally, K respondents are recruited for the rating task, with the kth respondent randomly shown some political statement S_i. Each respondent makes an inference about the statement, denoted here as $R_{ik} = 0, 1$. From random assignment, we can consider this inference to be a random variable, $R_{ik} \sim Bernoulli(\pi_{ik})$. The goal is to elicit many ratings, sampling from R_{ik}, to identify some *ad-level* ideology score

[6]A clear advantage to automated approaches is the ability to explore a high dimensional space of word frequencies, regardless of whether obscure language is used. Crowdsourcing the analysis of bill sponsorship, for example, would likely be ineffective (or very expensive) since legislative language is often lengthy and inscrutable.

$\mu_i = \Phi^{-1}(\pi_i)$, with Φ binomial link function.[7]

Clearly, repeated samples from R_{ik} would not be independent, since respondents are likely to remember their previous evaluations of the same statements. Instead, each ad S_i is randomly assigned, without replacement, to a subset of M respondents to produce a vector of 'perception' scores, $\tilde{R}_i = R_{i1}, R_{i2}, ..., R_{iM}$. Random assignment ensures that ratings of the same ad across different coders are independent, or $R_{ik} \perp\!\!\!\perp R_{im}$, $\forall k \neq m$, denoted here as *Scorer Independence* (A1):

$$E[R_{im}|R_{i1}, R_{i2}, ..., R_{ik_{k\neq m}}] \quad = \quad E[R_{im}]. \tag{3}$$

The number of M ratings per statement must be large enough to consistently measure π_i. In this design, N different ads (typically 8) are shown to each respondent, so that every ad-statement receives more than 140 ratings. This assignment produces a series of responses for each m coder, $\tilde{R}_m = R_{1m}, R_{2m}, ..., R_{Nm}$. Responses are also assumed to be independent of the order in which ads are rated, or *Item Independence* (A2):

$$E[R_{im}|R_{1m}, R_{2m}, ..., R_{n_{n\neq i}m}] \quad = \quad E[R_{im}]. \tag{4}$$

Order might matter if people learn over the course of the survey task. Yet, this will not systematically impact overall ad-scores when the number of ratings per respondent is fixed, since each ad is equally likely to appear at any point in the ordering.[8]

In this framework, the 'true' scoring μ_i of S_i can only be indirectly estimated by tapping people's perceptions, denoted as $\mu_{im} = f_m(\mu_i)$. Without additional assumptions, this inferential design can measure the degree of partisanship a particular sample of M

[7]The design is related to classic item-response testing (IRT), where an indicator $C_{ik} = 1 - (P_i - R_{ik})$, measuring whether an inference is correct, is modeled as a function of test-taker ability a_k, and item difficulty d_i: $p(C_{ik} = 1) = \Phi(a_k - d_i)$.

[8]The Appendix presents experiments showing that any possible item dependence is ignorable in practice, not surprising since there are relatively few coding tasks per respondent.

voters perceives about a set of ad-statements:

$$\begin{aligned} E[E[R_{i1}], E[R_{i2}], ..., E[R_{iM}]] &= E[E[\tilde{R}_i]] \\ E[\pi_{i1}, \pi_{i2}, ..., \pi_{iM}] &= E[\tilde{R}_i] \\ E[\pi_{im}] &= E[\tilde{R}_i] \\ \Phi(E[f_m(\mu_i)]) &= E[\tilde{R}_i] \\ \hat{f}_m(\mu_i) &= \Phi^{-1}\left(\sum_m \frac{\tilde{R}_i}{M}\right). \end{aligned} \tag{5}$$

If the sample is representative, then it is reasonable to interpret the score $\hat{\tilde{R}}_i = \sum_m \frac{\tilde{R}_i}{M}$, as an estimate of the population perception of partisanship in S_i. Initially, this provides an intuitive measure of how likely it is that S_i is airing a 'Democratic' or 'Republican' message in the minds of voters. Though not an ideal point, this metric captures meaningful perceptions of partisanship in communication, directly relevant to how voters access and incorporate political information in partisan context. This is further bolstered by the fact that voters are the *intended* targets of these messages, and thus are expected to be capable of appropriately interpreting the information contained in them.

The core concern is that crowdsourced perceptions may be biased or error-prone, so that scorings depend on a particular sample of coders. Complex, long or difficult tasks could yield low-quality responses, or require expertise and knowledge to navigate. In such cases, variation in coder ability could impact precision in the scores.[9] Another issue is that respondents' characteristics could influence their perceptions of ads. In the U.S., many voters have strong attachments to one of the two parties through their party identification (PID). A task that has voters evaluate platforms, news stories or advertisements featuring the parties or their candidates could suffer significant bias from the motivated interests of co-partisans (Ansolabehere and Brady 1989; Budak et al. 2016).

The key innovation in the approach developed here is to tap the *inferences* people

[9]This would be a problem if coders purely guessed, so $\pi_{im} \approx 0.5$, for a wide range of μ_i.

make about speech as a way to collect natural measures of partisanship or ideology. By removing all identifying party and ideology references, subjects are placed behind a *partisan veil of ignorance*, subject to uncertainty in the task. Since only the policy statements in ads are ever seen by respondents, we can be assured that this is the information they are using to make judgements. Further, respondents must first infer party, before motivated reasoning can be activated. And in facing uncertainty, guessers are likely to mute their motivated responses to avoid negatively evaluating co-partisans. Finally, the party guessing task is clear and straightforward, with right and wrong answers, making it simple to implement and easy for respondents to follow.

Nevertheless, interpreting *party perception scores* as estimates of latent ad-locations μ_i, requires an identicality assumption. The weakest possible assumption is *Weak Monotonicity* (A3), or $f_m(\mu) \leq f_m(\mu + \epsilon) \implies \mu \leq \mu + \epsilon$, $\forall f_m$. This ensures that $\hat{\tilde{R}}_i$ retains ordinality in μ_i. A stronger assumption is *Conditional Agreement* (A4) in perceptions:

$$f_m(\mu)|X_m \quad = \quad f_k(\mu)|X_k, \forall m \neq k. \tag{6}$$

In words, after controlling for differences in all relevant coder characteristics X, everyone can agree on how liberal or conservative each ad is in expectation, over the full range of μ. Combining A3 and A4, any continuous scoring $\hat{\tilde{R}}_i|X$, for statements $i = 1, 2, ..., S$, is a consistent estimate of voter perceptions of campaign positioning $\Phi(f(\mu_i))$.[10]

3.1 Implementing Guessing in CCES and MTurk

For this study, the guessing experimental design was implemented in three survey waves. The first was conducted in the 2014 Cooperative Congressional Election Study

[10]If $\Phi(f(\mu_i))$ is a continuous probability measure, then $\hat{\tilde{R}}_i|X$ retains cardinality in $f(\mu)$, though not necessarily in μ.. If ads are meant to be interpreted by voters, then it is most appropriate to focus on estimating their perceptions in $f(\mu)$, since these ground theories of electoral behavior, rather than considering some noumenal ad-location μ.

(CCES), while the second and third waves were fielded using subjects recruited through Amazon Mechanical Turk (MTurk) in 2015 and 2016. The CCES study included a total of 1,800 respondents, and was conducted two weeks before the 2014 election, with a follow-up post-election survey held the week after the election.[11] The 2015 MTurk study (wave two) recruited an additional 4,853 respondents, randomly assigning them into one of 6 survey frames, to be completed online through Qualtrics. The CCES included a large number of pre-treatment controls among the battery of common content questions. Additionally, 17 of these covariates were also collected in the 2015 MTurk study to assess and correct any differences between MTurk and CCES samples.[12]

The typical procedure in the experiment works as follows. Respondents first see a short statement indicating they are to read a set of positive ad statements in their entirety, and then to asses the party of the candidate airing the ad. Respondents then see 4 randomly selected positive ads, and guess the party of the candidate being promoted in the statement. Respondents next see a similar statement to read each negative ad presented, but now are instructed to guess the party of the candidate *being attacked* in the ad message. Respondents then see 4 randomly selected negative ads, and guess the party of the candidate under attack. For each response, subjects have the option to choose 'Democratic', 'Not sure', or 'Republican'.[13] (A screenshot of the general experimental protocol is included in Figure I in the Appendix.)[14]

---

[11]The current CCES data uses the survey weights and matching created by YouGov. Using the full, unmatched and unweighted data does not change any results.

[12]Controls include: gender, race, age, education, income, turnout, registration, 2012 vote choice, news interest, party majority in the House, ideological placements of the Democrats, Republicans, and oneself, and 7-point PID.

[13]The outcome ordering was randomly reversed *for each respondent*, with 'Not sure' always appearing in the middle. Thus, the ordering was consistent for each respondent, but randomly flipped across respondents.

[14]The pre-election CCES had 1,800 respondents assigned 8 ads each from a list of 100,

The 2015 MTurk survey extends and validates the party guessing results from the CCES, using 50 of the same ads included in that study, and 50 additional ads originally left out. Respondents in the MTurk *Frame 1* ($N = 1,268$) participated in an identical experimental frame as those in the CCES study. These respondents were asked to read 8 short positive and negative ads, and to guess the party of the candidate featured in each, using the same response, outcome and randomization structure described above. The structure in *Frame 2* ($N = 1,220$) is identical to that in *Frame 1*, with the important exception that respondents were asked to infer the ideology, rather than party of featured candidates. Here respondents choose whether each ad promotes or attacks a 'Liberal' or 'Conservative' candidate, or that they are 'Not sure', with these similarly randomly reversed. The remaining *Frames 3 - 6* provide robustness checks on the basic design.[15]

In the first two waves, 200 ads were chosen from 3,255 ads aired in the 2008 House and Senate elections as collected by the Wisconsin Ads Project (CMAG). The ads were sampled to balance a number of important factors. First, the ads exactly balance partisanship, with 100 Democratic and 100 Republican ads chosen, split evenly amongst positive and negative ads. ('Contrast' ads are excluded here.) The ads were also chosen

yielding $E[M] = 144$ expected guesses per ad. In the post-survey, 1,000 respondents were assigned 10 ads from a separate list of 50. This ensured $E[M] = 160$ given the roughly 80% attrition typically found in the post-election CCES.

[15] *Frame 3* ($N = 1,060$) randomizes whether positive or negative ads appear first, to assess possible learning effects. *Frame 4* ($N = 922$) uses monetary rewards (+\$0.20) for correct answers and penalties (-\$0.20) for incorrect ones to encourage attentiveness and discourage overconfidence. *Frame 5* ($N = 648$) transposes positive ads to be negative, and the reverse, to assess if tone itself influences voter perceptions independent of content. *Frame 6* ($N = 533$), asks respondents to rate how *specific* ('Very Specific' to 'Very Unspecific') the policy information is in ads. Some respondents in *Frame 3* overlap with *Frames 1, 4* and *6*. Results and implementation are discussed in the Appendix.

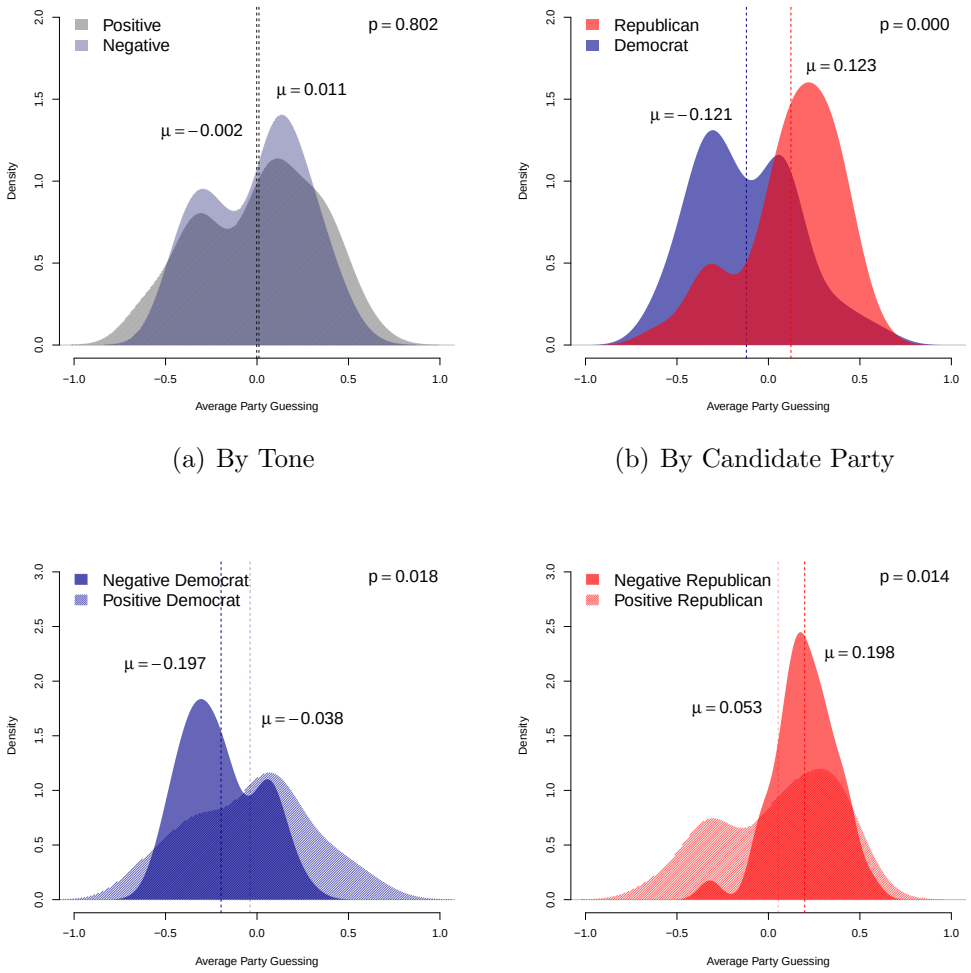
from among only those with at least some issue content as coded by CMAG, but were allowed to vary in how specific this policy information is, as well as whether it was accompanied by significant character or non-policy content. Further, ads were selected to maximize representativeness of the broader distribution of issues raised in the 2008 election. Finally, all 3,255 ads were scaled using *Wordfish* prior to the experiment, to ensure the 200 ads had significant variation on a latent text dimension. Overall, this balancing ensures that features of the statements (i.e., amount and type of issue content), do not correlate with ad tone in ways that might influence how respondents assess partisanship. This balancing is very successful, as measured by low intercorrelations exhibited between features of the included ads, especially by party and tone. Finally, ads were cleaned and processed, removing candidates' real names and partisan affiliations, and any other ideological or partisan terminology (e.g., liberal, conservative, centrist, bipartisan). Messages were then edited to be in the third person, and attributed to a generic candidate Clark.[16]

4 Dimensionality, Reliability and Validity in Party Guessing

This crowdsourced inferential framework can be used widely to measure a variety of dimensions in political texts, and marshaled to address many different questions in political science. The application here focuses on assessing how well the method captures partisan positioning in campaign advertisements as a critical test of the basic experimental framework. I first explore the dimensionality and reliability of the scoring method across the 2014 CCES and 2015 MTurk studies using the above 200 ads. In the next section, I examine the validity of these *party perception scores* through a series of additional tests.

The resulting *party perception scores* recovered for 150 ads in the CCES experiments

[16]The name Clark ensures a common baseline across experiments, though voters may use candidate gender to infer party (Goggin et al. 2015). In a third MTurk wave, I randomize name and gender, and find that male candidates only shift guesses additively rightward.



(a) By Tone

(b) By Candidate Party

(c) Democrats By Tone

(d) Republicans By Tone

Figure 2: **Density Plots of *Party Perception Scores* by Candidate Party and Ad Tone in the CCES**: Party perception scores, arrayed along the x-axis in the densities, are averaged over M guesses for 150 ads in the CCES sample, stratified by tone and party. Values of -1 indicate all Democratic and +1 all Republican guesses. Distribution means (μ) are indicated by vertical lines, and mean differences using t-test p-values.

are presented in Figure 2. The figure displays densities of perception scores by positive (light grey) or negative (dark grey) tone in Figure 2(a), and broken down by candidates' Democratic (blue) or Republican (red) party affiliation in Figure 2(b). The x-axis in these densities indicates the *party perception score* averaged over each ad, with -1 representing ads receiving all Democratic inferences, and +1 all Republican inferences.[17]

[17]Attack ads are 'flipped' for presentational clarity, so negative values indicate the target

Initially, we see clear and meaningful variation across the scale (overall accuracy is 58%) indicating that respondents are not just guessing randomly. Interestingly in Figure 2(a), the distribution of scores for positive and negative ads are statistically indistinguishable ($p = 0.802$), and have a bimodal shape, likely reflecting differences in the way the parties candidates' campaign. Indeed, Figure 2(b) shows a clear divergence in party perception scores across parties, with average Democratic-sponsored ads being perceived as airing more 'Democratic' messages ($\mu = -0.121$), compared to average Republican-sponsored ads ($\mu = 0.123$). Respondents do perceive many ads to be consistently and correctly associated with each party, giving credence to the measurement task for uncovering partisanship in advertising, using a high-quality survey like the CCES.

Quite a few ads are incorrectly classified (42%), however, including nearly 8% of ads being misclassified by two-thirds of the respondents evaluating them. An interesting pattern forms when looking at how tone interacts with party to drive misclassification. Figure 2(c) displays party perception scores for positive (light blue) and negative (dark blue) Democratic ads, while Figure 2(d) presents scores for positive (light red) and negative (dark red) Republican messages. An important result emerging from the plots is that partisan perceptions differ significantly across ad tone *within party*. People are much better at identifying which party is airing negative (63%), rather than positive ads (53%), and this is largely symmetrical across the parties. Thus, there are many *positive* Republican ads that signal consistent 'Democratic' messages, and many Democratic ads that make traditionally 'Republican' appeals, but counter-stereotypical *attacks* are rarer.[18]

A possible explanation for this result is that positive ads may convey limited issue

is perceived to be a Republican (hence is a Democratic attack ad), while positive values indicate the target is perceived to be a Democrat (hence a Republican attack ad).

[18]Counter-stereotypical attacks are interesting – Democrats are sometimes attacked for voting to increase health care costs or cut Social Security, while Republicans are readily attacked for raising taxes or increasing deficit spending.

information, making it hard for voters to discern distinct policy positions, while negative ads are more clarifying (Stokes 1992; Tomz and Van Houweling 2009). Alternatively, candidates may emphasize counter-stereotypical issues or positions to portray themselves as relative moderates in their positive ads, while attacking their opponents as partisan extremists (Henderson 2015).[19] Importantly, the feasibility of the measure does not depend on either theoretical view. Rather it relies on voters interpreting actual campaign signals in a process akin to how they might learn about competing candidates in real elections. In contrast, automated approaches to scale campaign speech could have considerable difficulty recovering this dimensionality, since these fundamentally rely on political speech to be non-strategic, distinguishing rather than obscuring candidates' policy differences.

The inferential framework developed here produces reasonable variation in *party perception scores*. Moreover, these highlight potentially interesting communication strategies employed by candidates that may elude traditional automated approaches to scaling text. Next, I present evidence that these party perception scores map directly onto an ideological dimension arrayed along on a liberal-conservative scale. For this, I replicate the inferential experiments from above using ideological rather than partisan labels to evaluate 100 ads in the 2015 MTurk study in *Frame 2*, as described above. A scatterplot of the resulting scores are presented in Figure 3. The plot presents average party guesses from the MTurk *Frame 1* on the x-axis (exactly replicating the CCES frame using 1,268 MTurk respondents), and average ideology guesses in *Frame 2* on the y-axis (utilizing 1,220 coders). As seen, the correlation between the two scores is $\rho = 0.97$, and the

[19]Results in the Appendix strongly suggest that avoidance or ambiguity *cannot* explain misclassification differences across tone. The 200 spots were selected to ensure positive ads had similar amounts of issue information as negative ones. Experiments in MTurk *Frame 6* show that party is misclassified at higher rates for more, rather than less *policy-specific* positive ads, with the opposite holding for negativity. And this difference is exacerbated when comparing similarly rated, specific positive and negative ads.

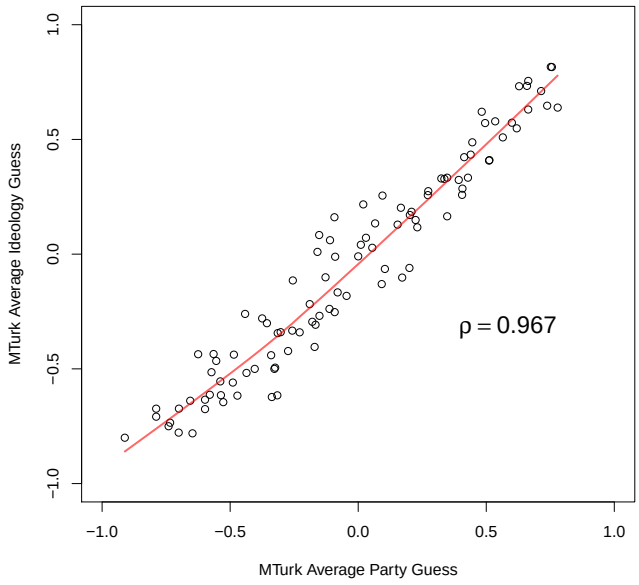


Figure 3: **Scatterplot of Perception Scores Inferring Party and Ideology**: Scores are averages of party (x-axis) or ideology (y-axis) inferences for 100 ads evaluated in *Frame 1* and *Frame 2* of the 2015 MTurk sample.

distributions are indistinguishable from each other. This finding strongly indicates that the likelihood an ad conveys a 'Democratic' message is identical to the likelihood it is communicating a 'liberal' one in the minds of voters. Consequently, party guessing also captures meaningful ideological information in ads.

Party perception scores are not interpretable as ideal points in a traditional way, since these are not based on any utility model of choice. One concern in interpreting this invariance between partisan and ideological perceptions, is that it could result from voters first inferring party in the ad, and then attaching some ideological label to it, simply from knowing that Democrats are liberal and Republicans are conservative.[20] We cannot peer into the minds of voters to see how they independently perceive party and ideology in

[20]The opposite inference from ideology to party is also possible. An implication of this two-step process might be greater variance in guessing ideology, since it involves making two different inferences about partisan ads. This greater variance does not emerge.

ads. Nonetheless, this evidence strongly demonstrates that ideology and party inferences produce indistinguishable measures of ad content.[21]

4.1 Assessing the Reliability of Party Perception Scores

As previously discussed, an important concern in using voters to evaluate political texts is that people might differ in how they rate the same statements based on certain of their characteristics (Ansolabehere and Brady 1989). This possible heterogeneity in voter inferences, encapsulated in A4 above (Eq. 6), could yield variation in aggregate party perception scores when crowdsourced samples differ significantly in their characteristics. Controlling for X covariate differences across coder samples could be necessary to ensure the inferential task is reliable, and that resulting perception scores are comparable. Any differences in PID seem particularly paramount given the central role it plays as a perceptual screen of information for voters in the American context (Zaller 1992).

I present the results here from a replication of the party guessing experiments using a convenience sample of MTurk workers in *Frame 1* to provide a strong test of the (conditional) reliability of the inferential approach to scale texts. Berinsky et al. (2012) explores the use of MTurk for survey research and finds that its population of workers who complete surveys are much less representative of the U.S. than samples recruited online by commercial vendors. I also find large and statistically significant differences on 16 of 17 demographic and political covariates collected in the both CCES study and the MTurk replication. (Differences are presented in Table I in the Appendix.) MTurk respondents are much younger, more likely to be male and white, less participating, and poorer, but better educated, more knowledgeable about politics, and most importantly, more Democratic and liberal. While all of these differences could have impact, the last two (Democratic PID and liberalism) could particularly skew party guessing in ways that

[21]Notably, this is also true of legislative ideal points which cannot distinguish between whether they measure partisan consistency or ideological extremity.

Table 1: **Effects of Individual Characteristics on Ad Guessing Outcomes**

	2014 CCES		2015 MTurk	
	Party Guess	Correct Guess	Party Guess	Correct Guess
Age	0.001	0.000	−0.000	0.000
	(0.001)	(0.000)	(0.001)	(0.001)
Female	−0.023	−0.006	−0.037	−0.025
	(0.016)	(0.013)	(0.021)$^{+}$	(0.020)
Black	−0.094	−0.081	−0.137	−0.064
	(0.028)***	(0.022)***	(0.042)**	(0.045)
Hispanic	−0.055	−0.041	−0.057	−0.004
	(0.029)$^{+}$	(0.024)$^{+}$	(0.046)	(0.049)
Registered	−0.014	0.009	0.043	−0.046
	(0.030)	(0.021)	(0.038)	(0.035)
Turnout	0.050	−0.012	−0.015	0.015
	(0.023)*	(0.018)	(0.025)	(0.023)
Education	0.009	0.018***	−0.005	0.003
	(0.006)	(0.005)	(0.009)	(0.008)
Income	−0.001	0.003	−0.001	−0.001
	(0.002)	(0.002)	(0.003)	(0.003)
News Interest	−0.011	0.010	−0.024	−0.046
	(0.008)	(0.007)	(0.014)$^{+}$	(0.013)***
Know House Majority	0.047	0.050	0.065	0.074
	(0.012)***	(0.010)***	(0.029)*	(0.028)**
Correct Party Placement	−0.008	0.024	0.028	0.148
	(0.013)	(0.010)*	(0.046)	(0.037)***
Self Placement	0.010	−0.001	−0.006	−0.001
	(0.010)	(0.008)	(0.016)	(0.013)
Party Identification	0.044	−0.001	0.044	−0.007
	(0.006)***	(0.006)***	(0.012)	(0.011)
Party Identification (Absolute)	−0.013	−0.006	0.016	−0.000
	(0.021)	(0.016)	(0.012)	(0.011)
Presidential Vote	0.076	−0.026	0.010	−0.004
	(0.015)***	(0.013)*	(0.013)	(0.012)
Observations	22617	22617	10144	10144
Survey Clusters	1797	1797	1268	1268
Ad Clusters	150	150	100	100
R^2	0.047	0.012	0.015	0.011

All models are OLS, and include additional controls, with unit and ad cluster standard errors.
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{+}$ $p < 0.1$.

bias raw comparisons across the two sample frames.

The scope of these concerns, *a priori*, can be seen in Table 1. The table presents the results from a regression of individual party inferences on important demographic and

political variables for both the CCES and MTurk studies.[22] Coefficients in columns 1 and 3 show the impact each characteristic has on the *direction* of party guessing, while those in columns 2 and 4 show effects on *correct* inferences. A number of findings stand out. People with more education or political knowledge make fewer errors in inferring an ads' partisan source. Yet in both studies, women, black and Hispanic voters are all more likely to infer an ad is Democratic, as are Democratic identifiers and those voting for President Obama in 2012. Conversely, white men, Republicans and Romney supporters are more likely to infer an ad is Republican. Not surprisingly, these characteristics also predict greater rates of misclassification. Reflecting a related finding reported in Ansolabehere and Brady (1989), respondents appear to believe that their own partisan or ideological complexion is mirrored in the ads they are randomly shown in the task. Party perception scores thus may indeed depend on the particular sample of coders used to rate ads.

Given these individual-level findings, I present a series of scatterplots comparing average party perception scores for 50 overlapping ads from the CCES and the MTurk *Frame 1*, with and without controlling for important covariates. Controlling for X was conducted through 'residualizing' the perception scores, retaining the average of the residuals for each ad from the regressions presented in Table 1, so scores only retain systematic variation orthogonal to sample covariates. Scatterplots are presented in Figure 4. Immediately we see in the plots that perception scores look essentially identical across the samples regardless of conditioning on X. The top-left panel shows raw scores for the CCES and MTurk, which are highly correlated at $\rho = 0.94$ *without adjusting for any covariates.* This invariance is remarkable given the above evidence about systematic factors influencing individual perceptions. Moreover, this reliability is retained across all four panels, with and without controls – if anything, residualizing out the influence of X slightly *reduces* the similarity in scores as shown in the bottom-right panel, though correlations never

[22]Regression coefficients are OLS, with standard errors clustered at both the respondent- and ad-level due to both having repeat evaluations in the task.

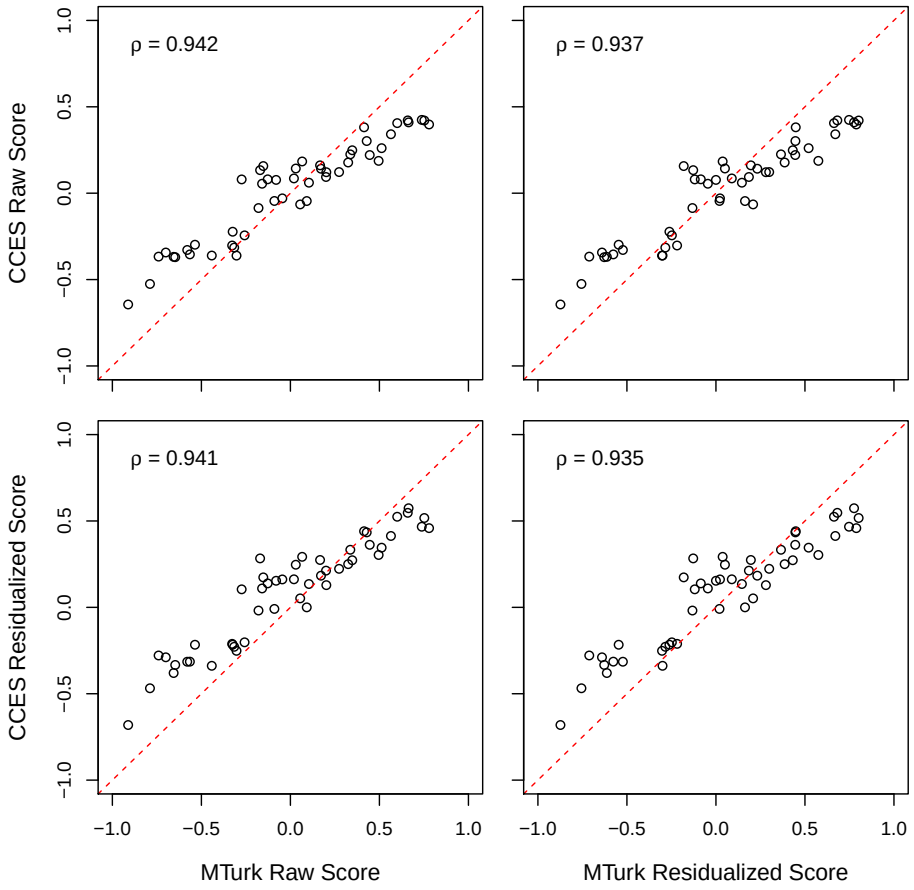


Figure 4: **Scatterplots Comparing CCES and MTurk *Party Perception Scores*, Before and After Conditioning on Covariates**: Data are 50 ads that overlap between the 2014 CCES and 2015 MTurk samples. Raw scores are the sample averages of ad party inferences, while residualized scores are averages of the residuals for each ad taken from an OLS regression of individual ad-inferences on X covariates, presented in Table 1.

drop below $\rho = 0.935$. One central explanation for this reliability is that by randomizing texts and placing voters behind a veil of uncertainty, the inferential task eliminates any systematic dependence between candidates' party and respondents' partisan evaluations, so the influence of individual-level characteristics cancels out in the aggregate.[23]

[23]Differences do emerge across frames. MTurkers answer 'Not sure' less often, increasing score variance. A small shift towards Democratic guessing for positive and Republican guessing for negative ads is seen, as MTurk has nearly twice as many Democratic than Republican identifiers, though the same rank-ordering is retained across both frames.

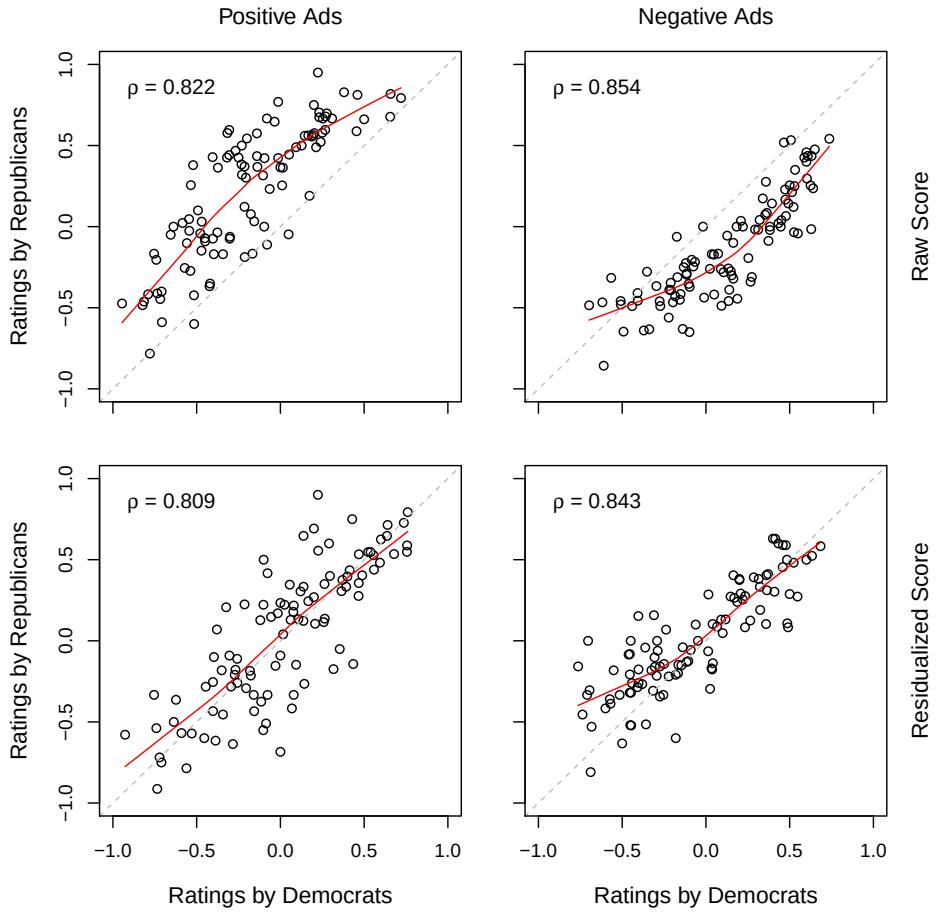


Figure 5: **Scatterplots Comparing *Party Perception Scores* Using Only Democratic or Republican Raters, Before and After Conditioning on Covariates**: Data are 200 ads scored by both the 2014 CCES and 2015 MTurk samples, stratified by respondent party identity (PID). PID is measured using a standard 7-point scale, with leaners coded as partisans. Raw scores are the sample averages of ad party inferences, while residualized scores are averages of the residuals for each ad taken from an OLS regression of individual ad-inferences on X covariates, presented in Table 1.

In a final test of reliability, I stratify party guesses by PID. This stratification can illuminate how much partisans differ in their inferences for in- versus out-party ads. These differences could pose particular challenges if identifiers are significantly imbalanced across coder samples. Alternatively, if bias in partisan guessing is simply additive, and partisans agree on the ordering of ads, then this would be a minor issue in scaling. Figure 5, presents scatterplots comparing average ad-scores as provided just by Democratic or Republican identifiers. Responses are combined for both MTurk and CCES

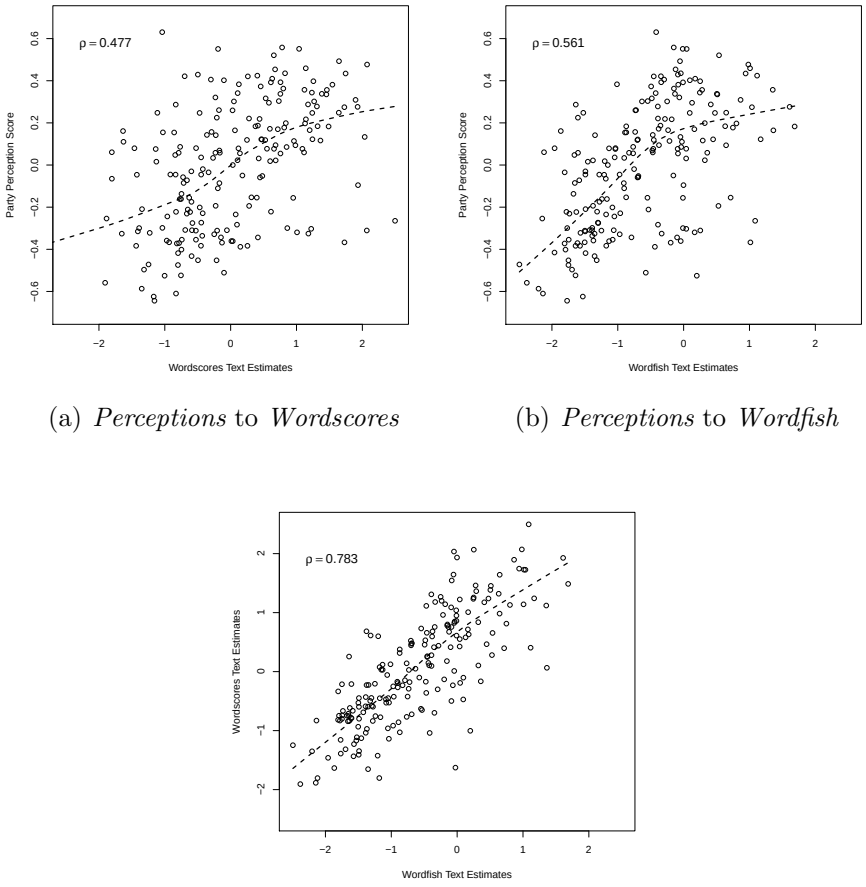
samples over the full set of 200 ads. Ratings for positive ads are presented in the left two panels and negative ads in the right two. The top panels show raw scores, and the bottom displays scores after controlling for covariates, following the above. As shown in the top two panels of Figure 5, we see Democrats and Republicans are both more likely to believe that positive ads feature candidates from their own party (*proximity bias*), while negative ads attack the other party's candidates (*extremity bias*). Importantly, the bias in these distributions is essentially additive – controlling for respondent characteristics, including PID, virtually eliminates the bias, as seen in the bottom two panels. Throughout, stratified party perception scores are robustly correlated at $\rho > 0.8$, putting a ceiling on the maximum possible PID bias in the experiments.

This fundamental invariance, even when comparing guesses across party identifiers, is an important feature of this inferential approach to scaling texts. In aiming at the right answer, very different respondents are able to agree about what constitutes the best guess (on average) about party. Party guessing produces scores that reflect a remarkable agreement about the rank ordering of ads as scaled going from most Democratic to most Republican. And at least with PID, bias appears to be mostly additive, and thus is likely to be cancelled out in the aggregate. Other biases may emerge, but it is hard to imagine any more powerful in this context than PID. Further, the above invariance in overall guessing between MTurk and CCES adds weight to the claim that any such biases cancel out in the aggregate as well.

4.2 Validating Party Perception Scores

Party perception scores are invariant to whether subjects are asked to infer the party or ideology of candidates, or differ along a range of important covariates, like PID, education or political interest. These findings strongly suggest that through an inference task, very different voters can agree in the aggregate on the ordering of ads based on their partisan and ideological content. While consistent across survey contexts, however, it is

possible that voter perceptions may not be the ideal way to scale political texts. Coders could be collectively myopic, consistently misjudging the content of ads. Or exposure to the ad-texts could have alternative treatment effects that influence voters in ways that feedback into the rating task. Fundamentally then, it is important to validate this method of crowdsourcing scaling through a number of tests, and in particular by comparing perception scores against the best-practices alternatives, *Wordscores* and *Wordfish*.



(a) *Perceptions* to *Wordscores*

(b) *Perceptions* to *Wordfish*

(c) *Wordscores* to *Wordfish*

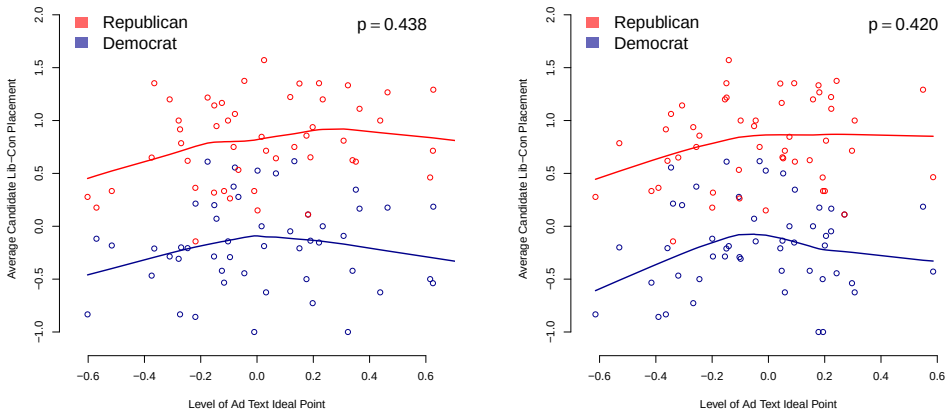
Figure 6: **Scatterplots of *Party Perceptions Scores* Compared to Both *Wordscores* and *Wordfish* Measures for 200 Ads**: Data are the sample of 200 ads. *Party perceptions* combine scores from both 2014 CCES and 2015 MTurk samples. For automated methods, ads are parsed into 1-gram words, stemmed and pruned of stop-words, then scaled using *Wordfish* and *Wordscores* models. Dashed lines are smoothed lowess trends.

Initially, I compare party perceptions to analogous scores produced using the *Word-*

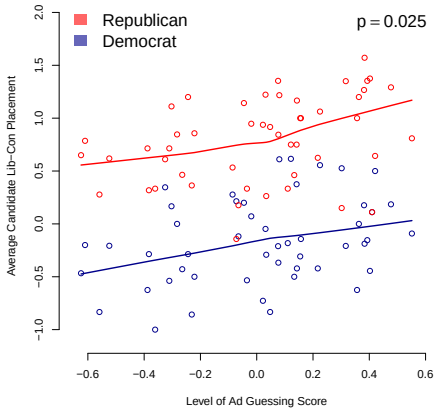
scores and *Wordfish* automated text-scaling approaches presented above in Equations 1 and 2. Text data are drawn from transcripts of the full list of 3,255 ads aired in the top-210 media markets in 2008 House and Senate elections as collected by CMAG. The transcripts are parsed into 1-gram words, then cleaned by stemming and removing stop-words. This cleaning process results in a matrix of 3,255 ads by 5,864 words, which are then scaled separately by positive (979), contrast (684), and negative (1,592) tone.

The results of this scaling analysis are presented in scatterplots in Figure 6 for the sample of 200 ads used in the above inference experiments. In Figures 6(a) and 6(b), *party perception scores* for the ads are displayed along the y-axis, with *Wordscores* or *Wordfish* scores displayed on the x-axis, respectively. As shown, party perceptions correlate positively with *Wordscores* at $\rho = 0.48$, and similarly with *Wordfish* at $\rho = 0.56$. Though scores correlate, there is considerable non-linear variation, especially as these approach their midpoints at zero. Hence text-based approaches produce related, but very different scales compared with party guessing. Indeed, this difference can be seen in Figure 6(c), which plots *Wordscores* and *Wordfish* scales, showing that each largely recovers the same dimensionality ($\rho = 0.78$) in word-usage. As described above, a potential explanation for this difference is that automated approaches may have a difficult time scaling ads when these are aimed at communicating strategic messages. Alternatively, it is possible that respondents miss important information in ads that automated approaches are better at measuring through models of the multidimensional space of words.

A way to assess these alternative interpretations is to identify whether text- or human-based approaches perform better in explaining the effects ad exposure may have on voter behavior or attitudes. For this I implement a series of vignette experiments in the CCES. In these, I ask respondents to assess two real candidates running in 2014, given their policy positions and a randomly selected ad statement attributed to one of the candidates. A representative protocol for the vignette is presented in Figure II. The candidates evaluated



(a) *Wordscores* (b) *Wordfish*



(c) *Party Perceptions*

Figure 7: **Influence of Ad Exposure on Candidate Placement, By Method of Scaling Ads**: Plots display the results from a 2014 CCES vignette experiment. These show the effect of randomly assigning a Democratic (Tom Hill) or Republican (Mark Meadows) House candidate one of 50 positive ad statements on voter placements of the candidates on a 7-point ideological scale ('Very Liberal ' to 'Very Conservative'), given three different ways to score ads: *party perceptions*, or *Wordfish* and *Wordscores* automated methods. Trend lines are lowess, and p-values are OLS, pooling over both candidates.

in the experiment are Mark Meadows and Tom Hill.[24] In the protocol, respondents are shown a brief preface, which includes the positions each candidate took on four roll call votes, along with descriptions of the votes.[25] Below this policy information preface,

[24]Meadows and Hill competed against each other in North Carolina's 11th House district in 2014, heightening the realism of the experiment.

[25]The votes included were: Middle Class Tax Cut Act, Simpson-Bowles Budget, Tax Hike

respondents are then randomly shown an ad statement from one of the two candidates. The candidate who appears is randomly selected. The message that is then attributed to the selected candidate is also randomly chosen, drawn from the set of positive ads scored previously in the CCES party guessing experiments.[26] Respondents then indicate where they would place both candidates on a liberal-conservative scale, as well as whether they would support either of the candidates. The main finding of interest here is whether being randomly exposed to a candidate's ad influences voter impressions of that candidate in ways that are consistent with the ideological scores produced through party guessing or any of the text-based scaling alternatives.

The results of the vignette experiments are presented in Figure 7. The figure plots the average liberal-conservative placement given by respondents for each randomly selected candidate and each randomly shown ad statement. Average placements of Tom Hill (in blue) are more liberal (i.e., closer to -1 that +1) overall on the y-axis, than are average placements of Mark Meadows (in red). Naturally, this is due the baseline policy information conveyed through the roll call positions, which are creating clear separation

---

Prevention Act, and Paul Ryan Budget. These votes were chosen due to their salience in legislative and party politics, as well as their inclusion in the list of common content questions in the CCES. The latter allows a comparison between voter attitudes on these items and evaluations of the candidates later on. The issues were arrayed from left to right according to their policy (cutpoint) locations recovered using DW-NOMINATE for the 113th Congress (Poole and Rosenthal 1997).

[26] An important cross-over element is used across the CCES, so that respondents never provide guesses for any ads that they could see in this candidate vignette experiment. Further, partisan information is never revealed or primed in these vignettes. The policy prefaces are meant to convey that the candidates have taken polarized positions on taxes and budgets, which should help voters make inferences about the overall ideological positions these candidates take on other issues, and possibly their partisanship.

between the candidates on the ideological scale. The x-axis in these plots indicates the scaled location of each of the ads, as scored by (a) *Wordscores*, (b) *Wordfish*, and (c) *Party Perceptions*. Thus, the plots illustrate how the average candidate placements change for each candidate associated with each ad, as the scaled ads go from most liberal to most conservative using each scaling approach.

Quite interestingly, as seen in Figure 7(a) and Figure 7(b), neither of the text-based scales correlates ($p > 0.4$) with how respondents place candidates airing those ad statements. Based on this finding, we might conclude that the ideological information in ads has little impact on how voters rate candidates campaigning on these messages. Yet, these vignettes illustrate this might be a hasty conclusion. In Figure 7(c), we see that the ideological information in ads, as scaled by (other) coders, is a significant predictor of candidate placement ($p = 0.025$). Ads that are rated as more liberal (conservative) through party guessing, when randomly associated with candidates, correspond with more liberal (conservative) candidate placements. This pattern holds for *both* the relatively liberal (Democrat) Tom Hill and the relatively conservative (Republican) Mark Meadows.[27] At least in this survey context, voters do respond to candidates' policy statements by updating their perceptions in line with received messages. And, most importantly, the policy information that seems to matter most in these messages, appears to be best measured by aggregating other voters' perceptions about the ads. Of course, text-based scalings may perform better in other contexts. Yet, this evidence adds caution to the general use of automated approaches to analyze the effects of messages received by voters.

[27]Figure 7 just presents the bivariate correlations. These persist when adding both additional ad- and individual-level controls.

5 Extending the Scores to Assess Theories of Campaigning

Lastly, I expand the inferential design to scale a much larger sample of 2,103 positive, contrast and negative ads drawn from the full set collected by CMAG for 2008. The principle aim is to show how the use of this inferential scaling method can help clarify certain contested theories of congressional campaigning. Scholars disagree over whether candidates' messages reflect the attitudes of partisan or centrist voters in their districts or their prior legislative records (Abramowitz 2011; Ansolabehere et al. 2001; Downs 1957; Jacobson 2015; Jessee 2010; Henderson 2015; McGhee and Sides 2011), or alternatively avoid discussing issues and positions altogether (Stokes 1992; Tomz and Van Houweling 2009). There is a related debate over whether candidates reposition towards the political center or remain consistent over the course of the campaign (Clouse 2006; McCarty and Poole 1998; Tomz and Houweling 2014). By fundamentally improving the measurement of position-taking in text, *party perception scores* can shed new light on our theoretical understanding of how incumbents and challengers compete for legislative office in contemporary elections. An ancillary benefit of this larger collection is that it can further bolster the validity and reliability of *party perception scores*, by showing that these meaningfully predict behaviors of politicians in election mode, even when scores are collected using a non-random procedure in a highly non-representative coder sample.[28]

The 2,103 general election ads were scaled using 654 coders in a third MTurk wave implemented in 2016. In total, MTurk coders made 94,534 party inferences about these 2,103 ads on the basis of simulations using the prior MTurk and CCES scores. These simulations (presented in the Appendix), indicated that at least 40 inferences per ad would be needed to recover an expected correlation of $\rho = 0.98$ under repeated sampling. Additional simulations showed that approximately 60% of the 3,255 ads would need to

[28] As done in the Appendix, this expanded set of scores can also be incorporated into automated machine learning approaches to make useful predictions about political texts in other campaign years, and beyond campaign advertising.

be scored to make quality predictions about the remaining left-out portion. This latter finding motivated the choice to evaluate a random sample of 2,103 ads (64.6%), with the 1,152 remaining left out for further analysis.[29] Scoring in this third MTurk wave *did not* follow the above experimental protocol. MTurk workers were encouraged to perform as many coding tasks as they desired. Further, ads were *not* randomly assigned to coders, though the coding tasks were randomly ordered so these did not appear in systematically related ways (e.g., party or issue-focus of ads). Positive and negative ads were again evaluated separately. Half of the 518 contrast ads were randomly assigned to be evaluated alongside 1,165 positive ads, while the remaining half were evaluated with 420 negative ads.[30] Finally, each coding task involved making two inferences, one for each of a randomized pair of ad statements.[31]

In this experimental design, the modal respondent only provided two inferences (i.e., completed one paired-task), while the maximum number of inferences was 1,630 (815 paired-tasks), with a mean of 144.6 guesses (72.3 paired-tasks) per coder. Further, to maximize efficiency in the coding, *no covariates were collected* about this set of respondents, including information on their PID. Consequently and critically in this experiment, it was decided *ex ante* to preclude the possibility of using covariates to make any

[29]Notably, 55 of these latter left-out ads were also included in the original 200 sampled, so that supervised predictions using the 2,103 ads could be evaluated using actual party perception scores as benchmarks.

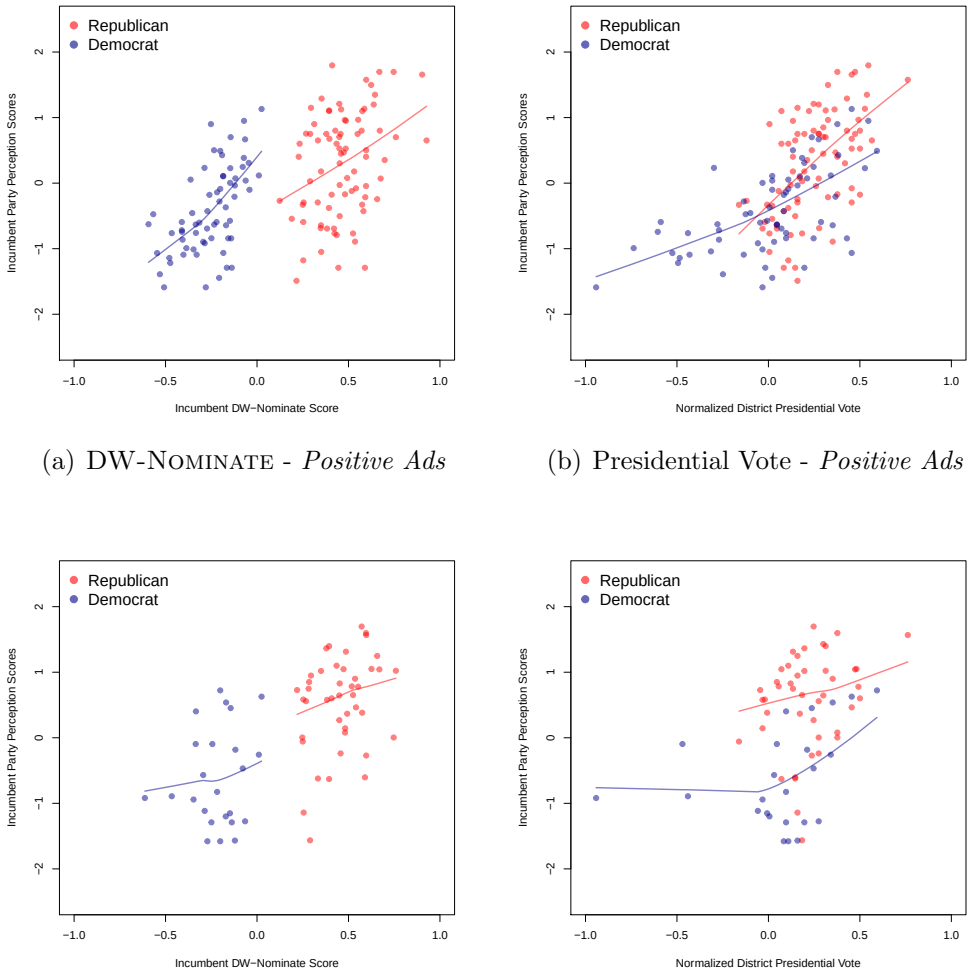
[30]This permits an analysis of the scoring effects of using a negative versus a positive frame for randomly assigned contrast ads, which typically include information about both attacked and promoted candidates. Interestingly, I find that attack frames induce more polarized ratings of contrast ads than do promotion frames.

[31]Candidates' names, including race and gender, were randomized to assess how these impact variation in inferences, especially interacting with issues. Candidates' states are also randomized when mentioned.

adjustments for sampling variation in the scores. Further, coders were encouraged to provide as many evaluations about ads as they wanted, potentially magnifying the non-representativeness of this as a convenience sample. In spite of these design choices, the resulting correlation between aggregate *party perception scores* for the 145 ads overlapping with the above sample, remained a robust $\rho = 0.90$, evidencing a remarkable reliability in the task. (A scatterplot of this association is shown in Figure XVIII in the Appendix.)

Turning to the analysis, following the above approach (shown in Figure 1), *party perception scores* for positive and negative ads are averaged for incumbents contesting elections in 2008, and compared with their prior DW-NOMINATE score, and the state- or district-level (normalized) presidential vote in their jurisdiction. These scatterplots are displayed in Figure 8, with the y-axis indicating average candidate perception scores, and the x-axis indicating the roll call or presidential voting measures. As shown in Figure 8(a), there is a clear, positive association between an incumbents' previous legislative record in Congress, and their scaled campaign *promotion* statements in the subsequent election. This association holds (roughly similarly) for both Democratic (blue) and Republican (red) incumbents. One obvious feature in this plot is that, while incumbents do campaign in ways that reflect their overall ideological positions in Congress, there is considerable divergence in how the two parties represent their districts when voting on roll calls. Yet, such a divergence is almost completely absent in incumbents' positive campaign positions. A partial explanation for this pattern may be found in Figure 8(b), which compares average campaign positions to voter attitudes. Average voters in states and districts are much more centrist in their attitudes, at least as measured by aggregate two-party presidential voting. Using party perception scores, we can see that congressional candidates intently target (centrist) voter attitudes in their election jurisdictions when running for reelection.

An altogether different pattern emerges when looking at positioning in negative ads. As displayed in Figures 8(c) and 8(d), we see a much weaker (within-party) association



(a) DW-NOMINATE - *Positive Ads*

(b) Presidential Vote - *Positive Ads*

(c) DW-NOMINATE - *Negative Ads*

(d) Presidential Vote - *Negative Ads*

Figure 8: **Correlation Between *Party Perception Scores* of 2008 Congressional Advertising and Roll Call Positions or Presidential Party Vote**: Data are transcripts of 2,103 House and Senate commercials for the 2008 election collected by the Wisconsin Ads Project (CMAG). Ads are scaled using a convenience sample from the 2016 MTurk study, with scores aggregated to the candidate-level, and standard-normalized. Roll call voting is scaled using standard DW-NOMINATE, and Presidential Vote is (normalized) state- or district-level proportion of party vote for president.

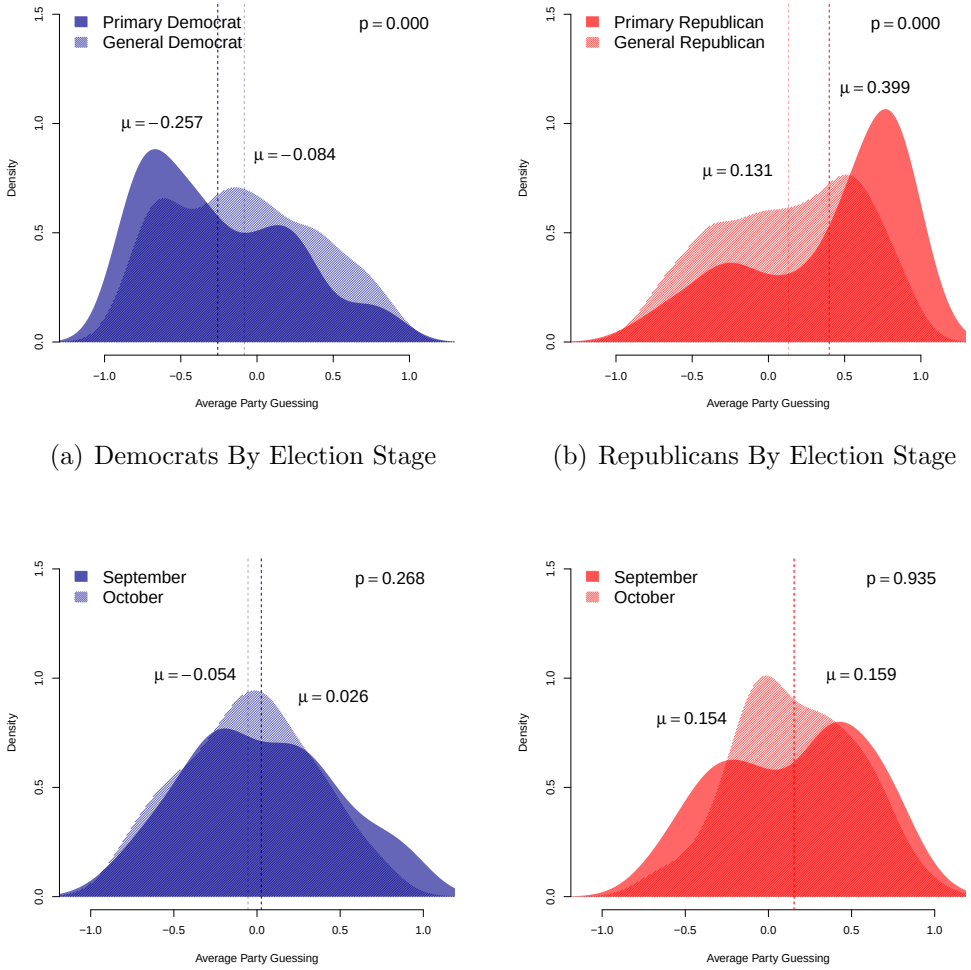
between the kinds of attacks incumbents air and their own prior legislative records or the attitudes of their constituents.[32] Yet, unlike in positive ads, position-taking in negative attacks *does* largely reflect the partisan polarization exhibited in Congress. This is seen

[32]There is a similarly weak association between a challenger's attack and the attacked incumbent's DW-NOMINATE score and jurisdiction presidential vote. Results are shown

here by the considerable divergence in the scores between Democrats and Republicans that is largely independent of a legislative record or district attitudes. Consequently, incumbents appear to be fighting to win the political center in a campaign, in an effort to differentiate themselves from their more partisan and polarized records in office, while pinning down opponents as truly out of touch with the voters (e.g., Henderson 2015). In broad strokes, using perception scores, we get a better glimpse of the contemporary campaign environment: candidates are somewhat faithful in discussing their legislative record during elections, but aim to mask, rather than clarify the issue priorities and positions that substantially differentiate them from their partisan opponents.

Given this tailoring, candidates may also target their ad messages to the mass of voters paying attention at particular moments in a campaign. Prior data limitations, however, have made it difficult to assess whether candidates substantially reposition over the course of an election cycle (Clouse 2006; McCarty and Poole 1998). Using new *party perception scores*, I explore the degree to which congressional candidates systematically drift during elections. Results are presented in Figure 9, which display densities of Democratic (blue) and Republican (red) candidates' positive ad-scores at various stages in the 2008 election. Figures 9(a) and 9(b) show scores just for Democratic and Republican challengers and incumbents who eventually make it to the general election. Not surprisingly, from these figures we see candidates clearly shift towards the political center as they move from competing in a primary to the general election. Interestingly, these also show that Republicans reposition further to the center (-0.268), than do Democrats (0.173), suggesting the former faced the more challenging primary and general election environment in 2008. Next, densities in Figures 9(c) and Figures 9(d) show scores for candidates airing at least one positive ad in both September and October. In contrast, these figures indicate that candidates largely remain ideologically fixed (on average) over the most active part of the campaign, the last two months. Democrats appear to shift slightly left from September

in the Appendix.



(a) Democrats By Election Stage

(b) Republicans By Election Stage

(c) Democrats By Election Month

(d) Republicans By Election Month

Figure 9: **Density Plots of *Party Perception Scores* by Stage of the Campaign**: Party perception scores, arrayed along the x-axis in the densities, are averaged over M guesses for 1,165 *positive* ads in the 2016 MTurk sample, stratified by party and stage in the campaign. Values of -1 indicate all Democratic and +1 all Republican guesses. Ads are included only for candidates winning a primary, or airing ads in both September and October. Distribution means (μ) are indicated by vertical lines, and mean differences using t-test p-values.

to October (-0.08), though this is statistically indistinguishable from zero. There is weak evidence that messages become more ideologically similar in the lead-in to Election Day, though again the distributions are statistically indistinguishable.

6 Conclusion

I develop an experimental approach to scale the partisan content of political ads. Accordingly, I randomly assign messages to survey subjects who are asked to infer the party and ideology of featured candidates. I find that these *party perception scores*, produced using a crowdsourcing design, are synonymous to an ideological dimension in voters' minds, remarkably reliable across sample and design conditions, and significantly out-perform standard automated approaches to scaling text through validation. I also demonstrate that this inferential design is extensible, and can reliably score a large number of ads (in a cost-effective manner), yielding much less measurement error than automated alternatives. This method represents an important advance in scaling text, particularly in contexts where strategic speech frustrates standard ideal point models.

This inferential method is likely to be most effective when speech is strategic or sparse in ways that limit the accuracy of text-based approaches. Even when automated scaling works well, inference experiments can be a useful mode of validation to insure that the dimensionality recovered is both grounded in realism and meaningful to voters. This latter point is especially important when validating scores, which necessitates making a theoretical assumption about the relationship between objects being scaled, and objects used in validation. Typically, any deviation between a new scale and its 'correct' benchmark is thought to indicate poor measurement, as opposed to imperfect theoretical knowledge. For example, there is heated disagreement about how to model position-taking in campaign competition. Using legislative voting to predict candidate advertising makes the assumption that ads are meant to reflect prior legislative positions. Yet, a plausible alternative is that politicians pander to centrist voters. Each assumption suggests a different way to validate ad-scores, either or both could be incorrect, *and* the only way to be sure is to validate both benchmarks against an otherwise valid scaling of ads. Such efforts could continue in endless circularity between measurement and theory.

Inferential measures offer a path out of this circularity. Party perceptions measure

whatever partisan information voters actually observe in ads. Hence this can provide an independent basis (e.g., replicability, exploring bias conditional on covariates) to judge the quality of the resulting scale. If campaign positioning is seen to reflect an appeal to centrists as scaled by voters, then this illuminates how average citizens understand the campaign messaging they are likely to receive. This does presume that ads are meant to be observed and understood by voters. But if false, then this probably precludes the scholarly enterprise of scaling ads. By tapping independent voter judgements, *party perception scores* can clarify important theoretical questions, without needing to reference other scalings to support its validity. Breaking the tautology, inferential scores can enhance the information value of comparing automated scalings to each other by providing evidence about the assumptions guiding those comparisons.

Beyond scaling partisanship in ads, the inferential approach outlined here can be generalized to measure a much wider array of dimensions contained in speech and text data. One future extension is to randomize candidates' names, alongside gender and race, to investigate how voters incorporate these cues when judging the ideological orientation of ads. Perception scores can also be incorporated into a wide variety of machine learning tasks, like measuring partisanship in other election cycles or in supervising the multidimensional scaling of ad-words. While the application in this study is scaling ads in the U.S. two-party context, the inferential design can be expanded to other kinds of speech, and in other partisan or political contexts (e.g., incumbency, candidate quality, multipartism). The limiting factor in extending this inferential approach is the cognitive complexity of the survey instruments, though to a lesser extent the crowdsourcing labor costs as well. Some tasks are not likely to be particularly well-suited to using voter inferences (e.g., regulatory policy, patronage party-systems). Yet, when both factors are low, crowdsourcing inferences may offer a powerful way to analyze text data, especially when prior automated approaches have proven unreliable.

More generally, there is no denying that the 'text-as-data' revolution has been im-

mensely productive in political science, and elsewhere. But this advance has not fundamentally eliminated or supplanted the importance of human judgement. Indeed, human coding should and is likely to play a critical role in validating automated analysis of text. Furthermore, in certain cases, human judgement can be a powerful and effective tool of measurement, that can compliment and improve the analysis of text data.

References

Abramowitz, Alan. 2011. *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy.* New Haven: Yale University Press.

Ansolabehere, Stephen and Henry Brady. 1989. "The Nature of Utility Functions in Mass Publics." *American Political Science Review* 83(1):143–164.

Ansolabehere, Stephen, James M. Snyder, Jr. and Charles Stewart, III. 2001. "Candidate Positioning in U.S. House Elections." *American Journal of Political Science* 45(1):136–159.

Beauchamp, Nick. 2010. "Text-Based Scaling of Legislatures: A Comparison of Methods with Applications to the US Senate and UK House of Commons." Accessed at `http://faculty.washington.edu/jwilker/tft/Beauchamp.pdf`.

Benoit, Kenneth, Drew Conway, Benjamin Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110(2):278–295.

Benoit, Kenneth and Michael Laver. 2007. "Benchmarks for Tex Analysis: A Response to Budge and Pennings." *Electoral Studies* 26(1):130–135.

Benoit, Kenneth, Michael Laver and Slava Mikhaylov. 2009. "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions." *American Journal of Political Science* 53(2):495–513.

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Budak, Ceren, Sharad Goel and Justin M. Rao. 2016. "Fair and Balanced? Quantify-

ing Media Bias through Crowdsourced Content Analysis." *Public Opinion Quarterly* 80:250–271.

Budge, Ian and Paul Pennings. 2007. "Do They Work? Validating Computerised Word Frequency Estimates Against Policy Series." *Electoral Studies* 26(1):121–129.

Clinton, Joshua D., Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 92(2):355–370.

Clouse, Clayton. 2006. "Candidate Ideological Shifting: Adapting to Different Electorates." Working Paper: `https://pantherfile.uwm.edu/ceclouse/papers/candidate-shifting.pdf`.

Downs, Anthony. 1957. *An Economic Theory of Democracy.* New York: Harper & Row.

Feinstein, Brian and Eric Schickler. 2008. "Platforms and Partners: The Civil Rights Realignment Reconsidered." *Studies in American Political Development* 22(1):1–31.

Gerring, John. 2001. *Party Ideologies in America, 1828 - 1996.* Cambridge, UK: Cambridge University Press.

Goggin, Stephen N., John A. Henderson and Alexander G. Theodoridis. 2015. "Party Guessed? Assessing Party Ownership of Issues and Traits with a Conjoint Classification Experiment." Presented at the Annual Meeting of the Midwest Political Science Association.

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):1–31.

Henderson, John A. 2015. "Distance in Advertising: How Candidates Use Issues to Distort Voter Perceptions and Influence Choices." Presented at the Annual Meeting of the Midwest Political Science Association.

Honaker, James, Michael Berkman, Chris Ojeda and Eric Plutzer. 2013. "Sorting Algorithms for Qualitative Data to Recover Latent Dimensions with Crowdsourced Judgments: Measuring State Policies for Welfare Eligibility under TANF." Access at `https://pages.shanti.virginia.edu/PolMeth/files/2013/07/Honaker.pdf`.

Jacobson, Gary. 2015. "Obama and Nationalized Electoral Politics in the 2014 Midterm." *Political Science Quarterly* 130(1):1–25.

Jessee, Stephen. 2010. "Voter Ideology and Candidate Positioning in the 2008 Presidential Election." *American Politics Research* 38(2):195–210.

Lauderdale, Benjamin and Alex Herzog. 2016. "Measuring Political Positions from Legislative Debate Texts." *Political Analysis* 24(2):374–394.

Lauderdale, Benjamin E. and Tom S. Clark. 2014. "Scaling Politically Meaningful Dimensions Using Texts and Votes." *American Journal of Political Science* 58(3):754–771.

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.

Laver, Michael, Kenneth Benoit and Slava Mikhaylov. 2011. "A New Expert Coding Methodology for Political Text." Accessed at: `https://www.princeton.edu/~pcglobal/conferences/methods/papers/Laver.pdf`.

Lo, James, Sven-Oliver Proksch and Jonathan B. Slapin. 2014. "Ideological Clarity in Multiparty Competition: A New Measure and Test Using Election Manifestos." *British Journal of Political Science* 44(1):1–20.

Lowe, Will. 2008. "Understanding Wordscores." *Political Analysis* 16(1):356–373.

Lowe, Will and Kenneth Benoit. 2013. "Validating Estimates of Latent Traits from

Textual Data Using Human Judgment as a Benchmark." *Political Analysis* 21(1):298–313.

McCarty, Nolan M. and Keith T. Poole. 1998. "An Empirical Spatial Model of Congressional Campaigns." *Political Analysis* 7(1):1–30.

McGhee, Eric and John Sides. 2011. "Do Campaigns Drive Partisan Turnout." *Political Behavior* 33(2):313–333.

Monroe, Burt L. and Ko Maeda. 2004. "Talk's Cheap: Text-Based Estimation of Rhetorical Ideal Points." Presented at the Annual Meeting of the Society for Political Methodology.

Monroe, Burt, Michael Colaresi and Kevin Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16(1):372–403.

Montgomery, Jacob M. and David Carlson. 2016. "Human Computation Scaling for Measuring Meaningful Latent Traits in Political Texts." Presented at the Annual Meeting of the Society for Political Methology.

Ororbia II, Alexander G., Yang Xu, Vito D'Orazio, and David Reitter. 2015. "Error-Correction and Aggregation in Crowd-Sourcing of Geopolitical Incident Information." *Proceedings of Social Computing, Behavioral Modeling and Prediction.*

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

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54(1):209–228.

Riker, William H. 1996. *The Strategy of Rhetoric: Campaigning for the American Constitution*. New Haven: Yale University Press.

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.

Snyder, Jr., James M. and Michael M. Ting. 2002. "An Informational Rationale for Political Parties." *American Journal of Political Science* 46(1).

Stokes, Donald E. 1992. Valence Politics. In *Electoral Politics*, ed. Dennis Kavanaugh. New York: Oxford University Press.

Tausanovitch, Chris and Christopher Warshaw. 2013. "Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities." *The Journal of Politics* 75:330–342.

Tomz, Michael and Robert P. Van Houweling. 2009. "The Electoral Implications of Candidate Ambiguity." *American Political Science Review* 103(1):83–98.

Tomz, Michael and Robert Van Houweling. 2014. "Political Repositioning: A Conjoint Analysis." Available at: `https://web.stanford.edu/~tomz/working/TomzVanHouweling-2014-08-15.pdf`.

Zaller, John. 1992. *The Nature and Origin of Mass Opinion*. Cambridge, UK: Cambridge University Press.