Using Experiments to Improve Ideal Point Estimation in Text with an Application to Political Ads*

John A. Henderson†

Assistant Professor

Political Science

Yale University

July 13, 2015

Abstract

Researchers are rapidly developing new automated techniques to scale political speech on an ideological dimension. Yet, the task has proven difficult across many settings. Political advertisements, in particular, have eluded such efforts. Candidates air relatively few ads, containing limited policy information, and there is little agreement about how to model political speech in the campaign, much less in general. Rather than model the underlying ideological structure of words, I develop an experimental approach to directly measure the content of political ads. I randomly assign ads to subjects, recruited in a large-N survey, who are asked to guess the party (or ideological leaning) of the featured candidate. Ads are then scaled as their expected partisan guessing score. This score is well-measured given random assignment and subject recruitment, and can be used in a supervised learning approach to scale other ad text. Due to the inferential nature of the task, subjects are less likely to exhibit bias in their guessing. Further, I show that the average partisan signal in ads is synonymous with an ideological dimension in the minds of respondents. I implement a number of tests to assess party guessing as a way to scale ads, each of which indicate remarkable reliability and validity in the approach. Finally, I explore ways to scale up the guessing task to a much larger set of ads. Beyond scaling ads, the inferential approach outlined here can be generalized to measure a much wider array of dimensions contained in speech and text data.

*For valuable comments I thank Devin Caughey, Jacob Hacker, Greg Huber, Stephen Goggin, Alex Theodoridis and Jonathan Wand. 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; Diermeier et al. 2011; Grimmer and Stewart 2013; Slapin and Proksch 2008). Though early pioneers in this research mostly used human coders to analyze text through content analysis, most recent work has turned to automated approaches to scale words and their usage on a common space (Laver et al. 2003; Slapin and Proksch 2008; Spirling 2012). Yet, scaling text data, and in particular political speech, has proven to be challenging. Our current models of political speech are at best rudimentary and disputed, and the standard utility accounts of legislative choice often do not transport well as heuristics for understanding word choice. Political speech is often less constrained than other forms of political behavior (e.g., floor voting with party agenda setters), and likely to be strategic, often obfuscating or avoiding issues. Compounding these challenges, sparsity in text data can make it difficult to model more complex uses of language, and especially the way ideological meanings may depend on how words interact with each other.[1]

Instead of modeling the underlying ideological structure of words in an automated fashion, I develop an experimental approach that taps human judgement to directly measure the content of political speech. In the design, I randomly assign ad statements to subjects, in a large-N survey context, and ask them to assess features of the ads. Specifically, I ask subjects to guess the party (or ideological leaning) of the candidate featured in each ad, on the basis of the text information contained in it. Ads are then scaled as their average partisan guessing score, which can be estimated (un)conditionally on survey covariates. Because of the inferential nature of the task, subjects are less likely to guess in ways that exhibit partisan bias. And any such bias can be corrected through

[1] Certain words may convey different ideological impressions depending on whether those words appear alongside or interact with any combination of all the rest in a corpus, e.g., "a woman's right to choose" vs. "a woman's right to choose her doctor."

covariate adjustment. I show that party guessing produces a scale indistinguishable from a liberal-conservative ideological dimension in the minds of voters, who score ads identically when evaluating either party or ideological labels. Though this scale does not represent ideal points in a conventional sense (which after all originate from utility models of choice), *party guessing scores* capture ideological and partisan information in ads that is meaningful and accessible to voters, something that fundamentally cannot be assured with ideal point estimation.

I implement this party guessing design using 1,800 respondents in the 2014 Cooperative Congressional Election Study (CCES) to scale 150 positive and negative congressional ads taken from the 2008 election. Subjects read 8 to 10 randomly assigned ad statements, guessing the party of featured candidates. Average party guesses are then compared to alternative ideal point estimates of ads using text, specifically *Wordfish* and *Wordscores* (Laver et al. 2003; Slapin and Proksch 2008). Each approach is then assessed through a number of tests. First, I replicate the party guessing experiment using subjects recruited through Amazon Mechanical Turk (MTurk), to scale 50 ads from the above 150, and an additional 50 ads left out of the original CCES study. Remarkably, party guessing using MTurk subjects produces virtually identical scores as that recovered in the CCES, *without adjusting for a single covariate*, indicating that the guessing task is invariant to substantial differences between subjects recruited across the sampling frames.

I conduct a second validation through a candidate vignette experiment in the CCES. In this experiment, real candidates from the 2014 election are presented to subjects, along with baseline information about their positions on four budget and tax roll call votes. One of the two candidates is randomly selected to make an additional statement in the vignette, with that statement being randomly selected from the set of 75 positive ads (of the original 150) featured in the guessing experimental frame. Subjects are then asked to place each candidate on a liberal-conservative scale, and to indicate which candidate

they would support in an election. Notable to the design, the randomly selected ads were previously scored through party guessing by *an entirely different sample of survey respondents*. A core finding from the vignette experiment is that the way subjects locate and choose among the two candidates significantly correlates with how other subjects previously scored partisan information in the ads. In contrast, *Wordfish* and *Wordscores* measures do not correlate with how respondents evaluate candidates making the ad statements, suggesting the party guessing scores are better measures of the ad information voters actually use when choosing between candidates airing them.

Finally, I assess whether party guessing scores reflect candidate targeting in the 2008 election. I link each ad to the House district in which it was aired, and then identify the correlation between the scaled ad positions and a normalized measure of district presidential vote. I compare this to analogous correlations using *Wordfish* and *Wordscores*. Party guessing scores are consistently (and sensibly) correlated with district presidential vote, while the text-based estimates are inconsistent predictors of district voting.[2]

Party guessing appears to outpace two of the most widely used automated approaches to scaling text. But, the cost of extending party guessing to a much larger number of ads may be prohibitive, making it worthwhile to accept some amount of measurement error for gains in scope and efficiency in scaling. A way to evaluate this trade off is to explore whether party guessing can be used in a supervised learning approach to make accurate predictions about other ads using the covariance between words and guesses. I do this in two stages. First, I assess how well a supervised learning approach (in this case the lasso-ridge elastic net) performs in predicting average ad guesses for the 50 held out ads using guesses from the CCES sample (Zou and Hastie 2005). The supervised learner predicts average scores that correlate with actual guessing at $\rho = 0.86$, indicating remarkable predictive accuracy just using ad words. This suggests economies can be obtained without needing respondents to guess every ad in a corpus.

[2]I replicate this for DW-NOMINATE and find the same pattern.

To map these economies of scale, I assess how accurate the supervised approach may be as the size of ad corpus increases. I develop simulations to identify the minimum (expected) number of guesses per ad that is needed to obtain a *mean squared error* less than a desired threshold. I then simulate the proportion of all ads that must be included to produce a supervised prediction that is expected to correlate with actual guesses at ρ for the remaining ads. I find that between 40 and 60 guesses per ad are needed to obtain sufficient convergence in the scores. I then show that a conservative bound on ρ is approximately linear in the proportion of ads to be guessed, so that at least 60% of the ads need to be included to produce a prediction that correlates at $\rho = 0.6$ with the guesses that would have been obtained for the other 40% left out. Consequently, researchers can reduce the scope of the guessing task by about half, and still recover predictions that are likely to outperform scalings obtained through automated IRT or word-scoring methods.[3]

The remainder of the paper outlines the guessing approach in greater detail, including a discussion of the particular implementation here to scale congressional ads. The paper concludes with a broader discussion of possible avenues to extend the basic inferential approach to augment the analysis of text data in general applications.

2 Prior Approaches to Scaling Speech

Scaling the attitudes and 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 research in this vein relied on experts or survey respondents to make evaluations about the liberal or conservative leanings of incumbents, parties or candidates. A standard way for this to be done involved having voters locate House members (or the parties) on some pre-defined ideological scale, usually ranging from 'Very Liberal' to 'Very

[3]This is based on the finding that ad guessing scores correlate with *Wordfish* and *Wordscores* at 0.54 and 0.45, respectively, using the 200 guessed ads in the CCES and MTurk experiments.

Conservative' (Ansolabehere and Brady 1989). Alternatively, researchers would rate incumbents based on their positions on select roll call votes defining the core ideological disagreements in Congress (Groseclose et al. 1999). These methods had significant shortcomings. Researchers found it difficult to choose an appropriate subset of key votes to rate members, and uncovered considerable bias in relying on partisan or low-information voters to evaluate co-partisan politicians.

An important innovation in this research was to move away from having researchers or voters make guiding judgements about the preferences of politicians, and to instead infer these from the large number of choices members make in office. This work is grounded in the study of roll call voting through a utility model of choice, with incumbents supporting proposals that are closer to their most preferred policy than the existing policies these aim to replace (Clinton et al. 2004; Poole and Rosenthal 1997). From this basic premise, using a high volume of legislative votes, researchers have produced a powerful measurement tool that can reliably summarize a great deal of the conflict within and across Congresses. Given this reliability, roll call measures have become important benchmarks to evaluate the myriad other approaches used to scale elite and voter attitudes.[4]

There is an interesting parallel in the way researchers have analyzed data taken from text sources. Again much of the early work employed experts or coders to glean meaning from the words or images presented in party platforms, floor speeches and campaign ads. A prominent line of this research, the 'Comparative Manifestos Project', uses a small number of experts to code the issues and positions contained in each sentence of hundreds of party platforms (Benoit et al. 2009). Other scholars have taken a more holistic approach, coding the unique policy positions taken overall in texts (Feinstein and Schickler 2008; Gerring 2001). Notably, this rubric approach has been the central way scholars have collected data on the content of campaigning and advertising (Riker 1996).

[4]Some of these include scaling campaign donations, twitter networks, legislative press releases, or elite survey responses.

An advantage to this method is that human coders can use their experience or judgement to interpret the policy direction of political messages, even when these are complex or ambiguous. Yet, these more traditional forms of guided content analysis have quickly been replaced by efforts to automate the granular analysis of text. This is mainly due to the explosion in the amount of political text able to be represented numerically, which has made it far easier to use machine-based approaches rather than expert coders to make inferences about large volumes of political communication.

Much of this automated scaling has explicitly extended the utility framework of legislative choice into the domain of political speech. Accordingly, scaling text takes place in an unsupervised manner, with choices over the use of particular words seen as driven by how 'close' those words are to describing a person's ideal policy position (Monroe and Maeda 2004; Slapin and Proksch 2008). One implementation of this is developed by Slapin and Proksch (2008) in their *Wordfish* model. (See the Appendix for a fuller elaboration of the model.) A way to represent this model is in terms of a word 'cutpoint', which defines the location in space where a legislator would be indifferent between choosing to utter a word and remaining silent. Though somewhat awkward, a cogent way to understand this cutpoint is to imagine that some words or phrases (e.g., "supporting a woman's right to choose") clarify a commitment to a particular ideological position, while silence meaningfully conveys a different policy commitment, and the cutpoint determines where legislators will be indifferent between the two.

Certainly this utility framework is an odd fit for modeling political speech. In that regard, it mainly operates as a practical heuristic, albeit one that has seen some criticism. For example, in the limit, certain theoretical accounts of advertising would argue that all words should have zero discrimination about ideological positions since speech is cheap talk, and candidates avoid talking about their policy positions (Stokes 1992; Tomz and Van Houweling 2009). Strategic speech more generally, if not appropriately modeled,

could induce serious bias in measurement (Monroe and Maeda 2004). Further, though the precise implications are unclear, there are deeper criticisms over how well speech can be modeled in terms of preferences, which emerge from axioms of choice and relations far removed from the nature of speech. Nevertheless, this unsupervised approach has proved successful in some contexts, and there are ongoing efforts to make improvements in the way speech is modeled in a political space (Kim et al. 2014).

An alternative to this unsupervised utility approach is supervised learning. One variant of this method is to build a dictionary of 'liberal' and 'conservative' words, and then score documents based on the frequency in which these ideological words are used (Beauchamp 2010; Laver et al. 2003; Lowe 2008). A common implementation of this dictionary approach works at the level of scoring words, hence taking the name *Wordscores* (Laver et al. 2003). (Again see the Appendix for a fuller elaboration.) Reference texts C and L are first chosen by the researcher to represent canonical conservative and liberal statements. Each W_j word is then scored as S_j based on the proportion of the time it appears in C rather than L, with $+1$ and -1 scoring weights attached to each proportion respectively. Then a document score is constructed by weighting the proportion of the document devoted to W_j by its score S_j, and averaging over all the weighted word proportions in the text.

Since their development, *Wordscores* have been both innovative and influential. Yet, these scores have some undesired properties. The approach assumes that words have equal discrimination weight (Lowe 2008), and for this and other reasons are likely to overfit the data (Lowe 2008; Monroe et al. 2008). It also eschews the use of prior information about words (or documents) that help smooth estimates when words exclusively appear in only conservative or liberal documents. In light of these issues, scholars have adapted this dictionary method to be more fully Bayesian, including developing a more elaborate (Bayesian) model of political language generation (Beauchamp 2010; Monroe et al. 2008).

A final alternative to this word scoring method is a more general approach to supervised learning. Here the target is to make a prediction about some annotation that is associated with documents as a function of the words that appear in them (Diermeier et al. 2011; Grimmer and Stewart 2013). A subset of d documents may have some pre-existing scoring of ideology $S^{(d)}$, for example, an indicator for party, or an ideological score taken from roll call voting. Regressing this annotation on a matrix of word counts can estimate the marginal influence each word has on variation in this outcome. If the number of words is large, and in particular, larger than the number of documents, some type of *regularization* is required to constrain the word parameters, and insure that the model can be identified. There are many ways to impose regularization, including the use of Bayesian priors on word coefficients, or through lasso and ridge regression, or their combination in the 'elastic net' (Zou and Hastie 2005). Though promising, this approach is rarely used to estimate ideal points in text data.

2.1 The Validation Tautology

Unlike unsupervised approaches, supervised prediction does not assume of model of ideological expression. The latter simply aims to make the best predictions possible on some outcome given the distribution of words across documents. In this sense, supervised methods are sometimes cast as efforts to refrain from making substantive assumptions about the way political speech is generated (Grimmer and Stewart 2013; Hopkins and King 2010). However, both of these automated methods fundamentally share a reliance on some form of validation to assess the quality of the resulting scale estimates (Benoit et al. 2012; Grimmer and Stewart 2013; Laver et al. 2011; Lowe and Benoit 2013). Further, this validation step *always* requires substantive judgements to be made by researchers, and these typically involve strong theoretical assumptions that cannot be tested. For example, a common validation strategy is to assume some prior ideological scale is *the*

correct benchmark that accurately depicts political behavior in some domain. The typical benchmark for legislatures is DW-NOMINATE or similar ideal point measures of legislative voting (Poole and Rosenthal 1997). When a new ideal point measure of legislators is developed, it invariably is compared to DW-NOMINATE to assess its correlation. This validation step is tautological in assuming that DW-NOMINATE is itself an accurate measure of legislative ideology. Scholars generally agree that DW-NOMINATE *is* an appropriate benchmark for other legislative scalings, which is not disputed here. But it does raise an interesting question about what exactly afforded this measure its benchmark status.

This strategy of validation extends beyond the legislative setting. Indeed, researchers scaling other kinds of political behavior (e.g., campaign donations, floor speeches, advertising) have also commonly resorted to showing how well their new scales correlate with DW-NOMINATE. The presumption is that high correlations between scales of roll call votes and of campaign effort suggest the latter is likely to be a valid measure of the way incumbents position in elections. Yet, here again lies a tautology, namely that roll call ideal points are the appropriate benchmarks to assess the behavior of candidates in the campaign. The fundamental problem is that a low correlation between a scale of campaign speech and another of congressional voting cannot tell us whether the text-based scale is poorly measured, or that it is wrong to assume that campaign positioning is meant to reflect roll call voting. Campaigning may in fact be about strategically appealing to more centrist voters in a general election, or alternatively to more extreme donors and primary supporters, rather than faithfully discussing votes taken in Congress. To verify which is the correct substantive assumption, however, we need independent information that the new scaling of speech is itself valid. But, this would obviate the need to validate automating scaling using a benchmark scale.

This validation stage may seem especially paramount when assessing unsupervised models of political speech that make strong assumptions about the way people communi-

cate (e.g., Lowe and Benoit 2013). Yet, supervised models double down on this tautology by assuming that whatever annotation is used for prediction is the right dimension to capture ideological positioning in campaign speech. In a dictionary approach, substantive judgement is used to decide which documents best represent the ideological poles. Often this is simply a subset of the Democratic and Republican texts in a corpus. This is the equivalent of assuming that party candidates aim to campaign faithfully on the basis of their partisanship (Beauchamp 2010). Using DW-Nominate as the annotation, alternatively, assumes that ads are meant to faithfully reflect legislative behavior (Diermeier et al. 2011). The best way around this tautology is to identify independent ways to validate text scaling (Benoit et al. 2012; Lowe and Benoit 2013). One possibility is to improve the modeling of ideological speech. Scholars could also advance our understanding of how exactly candidates position in a campaign given the records they compile in Congress. So far, progress on both fronts has moved slowly.

Rather than proceed in this way, I propose an altogether different approach, using human judgements to scale campaign text. When feasible, few methods are likely to improve on the direct use of voter ad evaluations. Since people are the targets of ads, we should reasonably expect that they are capable of comprehending the information contained in them (Benoit et al. 2012; Lowe and Benoit 2013). In comparison, unlike voters, machines do not possess the prior experience necessary to understand the context of campaign speech or to make valid judgements when rare or unusual words are used. The advantage in automated approaches is that machines can explore a high dimensional space of count indicators, an impossible task for humans. But, when this dimensional space is sparse, representing a large number of words that rarely co-occur, the advantage in automation is greatly reduced. (This is especially so when the amount of text in each document is short enough to be read in less than a minute.) Hence scales of text generated by humans, at least under certain conditions, are likely possess greater validity

than automated approaches. Also, this validity may be independent of how well these scales correlate with other measures, since the method is rooted in how people actually perceive the ideological information being analyzed. The next section discusses and develops this human coder approach in greater detail.

3 Using Partisan Inferences to Scale Ads

Using the collective judgements of survey respondents to scale ads fits within a broader, burgeoning framework of crowdsourcing coding tasks to collect and measure data. The general goal in crowdsourcing is to disaggregate specific tasks (e.g., rating short spans of text) and to distribute these widely to a large number of respondents to be evaluated in some way (Benoit et al. 2012; Budak et al. 2015; Honaker et al. 2013; Lowe and Benoit 2013; Ororbia II et al. 2015). The method harkens back to more traditional content analysis, with the important feature of collecting a large number of measurements for each item and each document. Much of this work so far has been dedicated to validating automated analysis or small-N coding (Benoit et al. 2012; Lowe and Benoit 2013), though increasingly is being used for data collection and analysis (Budak et al. 2015; Henderson 2015; Honaker et al. 2013; Ororbia II et al. 2015), including the measure of an ideological dimension (e.g., Benoit et al. 2012; Lowe and Benoit 2013).

In principle, crowdsourcing could be used for any number of coding tasks, and not just to scale ads or other documents. Honaker et al. (2013), for example, use a large number of pair-wise comparisons, asking respondents to rate whether a country is more democratic than another after reading a short description about each. Ororbia II et al. (2015) crowdsource an annotation task to have respondents determine if a story covers a militarized dispute. A possible concern in utilizing survey respondents in this way is that the quality of their judgements may be limited, biased and error-prone. When coding instruments are relatively clear, and the number of codings relatively small, crowdsourcing

is likely to perform well (Ororbia II et al. 2015). Yet, complex, long or difficult tasks are likely to elicit considerable measurement error, and low quality responses. Another problem is that respondents can be biased in their judgements. In the U.S. context, for example, many voters have strong attachments to one of the two parties through their party identification (PID). A task that asks voters to evaluate a platform, news story or advertisement involving one of the two parties or their candidates is likely to suffer bias due to the motivated interests of co-partisans (e.g., Budak et al. 2015).

The key innovation in the scaling approach developed here is to tap the *inferences* subjects make about text as a way to collect unbiased measures of partisanship or ideology. Due to motivated reasoning, having respondents directly rate ads that feature candidates from either party will likely produce scores that contain some amount of PID-drived measurement error.[5] By removing all identifying partisan and ideological references, subjects are placed in something of a partisan veil of ignorance when asked to infer the party of featured candidates. Since only the policy statements in ads are ever seen by respondents, we can be reasonably assured that this is the information they are using to make judgements. Additionally, the party guessing task is clear and straightforward, making it simple to implement and easy for respondents to follow. The task also has right and wrong answers, which may add additional motivation for respondents to try to guess correctly rather than expressively.[6] It is sill possible that respondents guess in partisan ways. Evidence presented below, however, shows this generally is not a concern.[7]

[5]Such a task might involve having respondents rate ads, that include party labels, going from 'Very Liberal' to 'Very Conservative'. It is possible that respondents will rate all out-party ads at the extremes of the scale, while rating in-party ads where they would locate themselves. This would produce more polarized ratings than otherwise expected simply due to expressive motivations.

[6]This inferential approach has some connection to the jury theorem. In the latter, if the probability of a correct guess for all respondents is $p_i(Correct) > 0.5$, and N people obtain independent and unbiased information about the outcome through the ad words, then their collective judgement will indicate the correct party. The aggregate guessing outcome will also converge on $E[p_i] = p$ as N increases. If $E[p_i] < 0.5$, then the 'jury' will collectively guess incorrectly about the party, with p essentially scoring the ad on how difficult or easy it is for subjects to guess correctly. One reason for this difficulty could be that Republican ads signal policy information that is commonly seen as Democratic, and the reverse.

[7]A reason to expect partisan bias to be low is that respondents must first infer party in an ad, before motivated reasoning can be activated. Since most respondents will be somewhat uncertain in

3.1 Implementing Guessing in CCES and MTurk

The above guessing design was implemented in two sets of experiments. The first was conducted in the 2014 Cooperative Congressional Election Study (CCES), and the second using subjects recruited in 2015 through Amazon Mechanical Turk (MTurk). The CCES study was conducted across two subsample modules.[8] In CCES *Module A*, 1,000 people were recruited and sampled in an online survey held two weeks before the 2014 election, with a follow-up post-election survey held the week after the election. In CCES *Module B*, 800 people were surveyed in a similar fashion. The MTurk study recruited 3,798 additional respondents, randomly assigning them into one of four survey frames (MTurk *Frame C, D, E* and *F*), to be completed online through Qualtrics.[9] Across both CCES modules, a large battery of common content questions were asked prior to the respondents being split into the subsamples, allowing for a large number of pre-treatment controls. Additionally, 14 of these covariates are also included in the MTurk study to assess and correct for any differences between MTurk and CCES recruitment.[10]

The general procedure in the experiment works as follows. Respondents first see a short statement informing them that 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. Next, respondents see a similar statement to read each negative ad presented, but now are instructed to guess the party of the candidate *being attacked*

their guesses, it is possible that their motivated responses will be muted in desiring to avoid the risk of negatively evaluating a co-partisan politician. Also, paying subjects for correct guesses has no effect on how ads are guessed overall, indicating, among other things, that any partisan bias is likely to be negligible in the baseline task.

[8]The current version of the data use the survey weights and matching created by YouGov, though future analyses will use the full, unmatched and unweighted data.

[9]The MTurk study assigned subjects into frames as follows: *Frame C* (1,247), *Frame D* (1,227), *Frame E* (644), and *Frame F* (653).

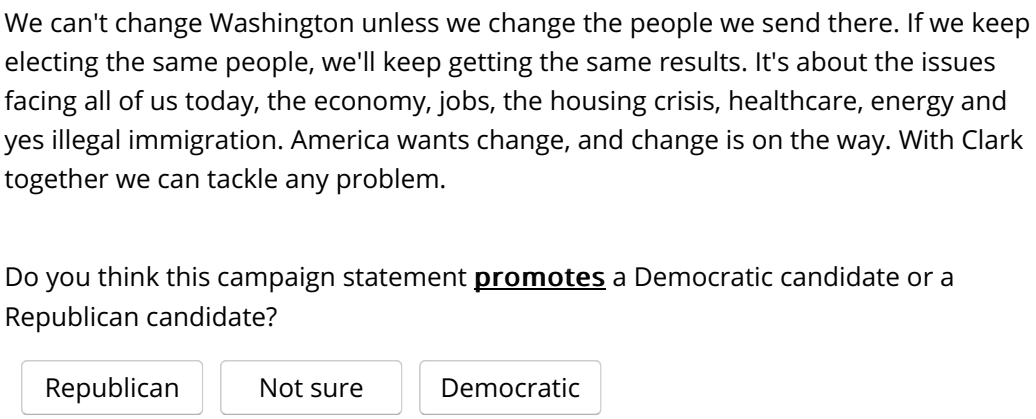
[10]These controls include: gender, race, age, education, income, turnout, registration, 2012 vote choice, news interest, party majority in the House, ideological placements of the Democrats, Republicans, and oneself, and 7-point PID.

in the ad message. Respondents then see 4 randomly selected negative ads, and guess the party of the candidate being attacked. For each guess, respondents have the option to choose either 'Democratic', 'Not sure', or 'Republican'. The order of these outcomes were randomly reversed *for each respondent*, with 'Not sure' always appearing in the middle. Thus, this outcome ordering was consistent for each respondent, but randomly reversed across respondents (sometimes 'Democratic' and sometimes 'Republican' appearing first). A screenshot of the general experimental protocol is included in Figure 1.

Respondents in both CCES *Module A* and *Module B* participated in an initial party guessing experiment in the 2014 pre-election survey. This survey randomly selected from a total of 50 positive and 50 negative ads, without overlap in ads between modules. Respondents in *Module A* then repeated the party guessing experiment in the post-election survey to assess an additional 25 positive and 25 negative ads. These experiments were identical to those in the pre-election survey, with the exception that each respondent was shown 5 positive and 5 negative ads. This was done in order to insure that each ad received a similar number of (expected) guesses given the roughly 80% attrition typically found in the CCES post-election survey. Thus, across both the pre- and post-election CCES experiments, a total of 150 ads were assessed by respondents using the above guessing frame to scale the partisan content of issues in ads.

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 C* 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 D* is identical to that in *Frame C*, with the important exception that respondents were asked to guess the ideology, rather than party of featured candidates. Here respondents

Figure 1: Experiment Protocol for Guessing Party from Positive and Negative Ads



(a) Positive Ad Guess

(b) Negative Ad Guess

choose whether each ad promotes or attacks a 'Liberal' or 'Conservative' candidate, or that they are 'Not sure'. The order of these follow an identical pattern as above, with outcomes randomly reversed and 'Not sure' always appearing in the middle.

MTurk *Frame E* returns to party guessing. This experiment aims to explain why positive and negative ads appear to signal different party information to voters. The approach is to transcribe positive ads into negative ones, and the reverse, before guessing.

Finally, *Frame F* evaluates what effect a financial reward for correct responses has on the way positive and negative ads are scaled through party guessing. Here MTurk subjects are provided an additional \$0.2 for each correct guess. Similar to the above, both *Frame E* and *Frame F* ask respondents to guess 8 positive and negative ads. Due to their smaller sample size, these frames randomly draw from a list of 50 total ads, 25 of which come from those used in both the CCES and above MTurk samples, and 25 of which are exclusive to the MTurk study. (The results from these frames are not central to the analysis below, and thus are discussed mainly in the Appendix.)

In total, 200 ads were chosen from 1,662 ads aired in the 2008 House and Senate general elections as collected by the Wisconsin Ads Project (CMAG). The ads were chosen 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.) The ads were also chosen based on having 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 their representativeness of the broader distribution of issues raised in campaign advertisements. Finally, prior to the experiment, the ad text was scaled using *Wordfish* (Slapin and Proksch 2008). The 200 ads were also chosen to insure significant spread on this text-ideological dimension.

Overall, this balancing insures that features of the ads do not correlate in ways that might influence how respondents assess partisanship. From a substantive perspective, it could be important to insure that negative and positive ads do not significantly differ in ways, unrelated to tone, that might make it easier for voters to infer party.[11] From a measurement perspective, this balance will help insure that any average bias in guessing will be 'balanced' or similar across many features of the ads. This balancing effort is

[11]In this sample, positive and negative ads score similarly on levels of policy and issue specificity as measured by CMAG. My own inspection of ads generally confirms this. Future effort will be devoted to highlighting the specific positions taken across the ads.

very successful, as measured by low intercorrelations exhibited between features of the included ads, and especially ad tone.

Finally, once selected, ads were lightly cleaned and processed. This involved removing candidates' real names and partisan affiliations, as well as any other ideological or partisan terminology (e.g., liberal, conservative, centrist, bipartisan). The messages were then edited to be in the third person, and were attributed to a generic candidate Clark.[12] The results from the party 'guesses' experiments can be used widely. Here these form measurements of the 'partisanship' or partisan ideology conveyed in the ads as perceived by voters. Moreover, these guesses can be studied in a variety of ways, including assessing the degree to which various features of ads or characteristics of voters improve the probability of correct inferences. Additionally, I use these inferences in an second embedded experiment below examining how ads influence voter impressions of candidates.

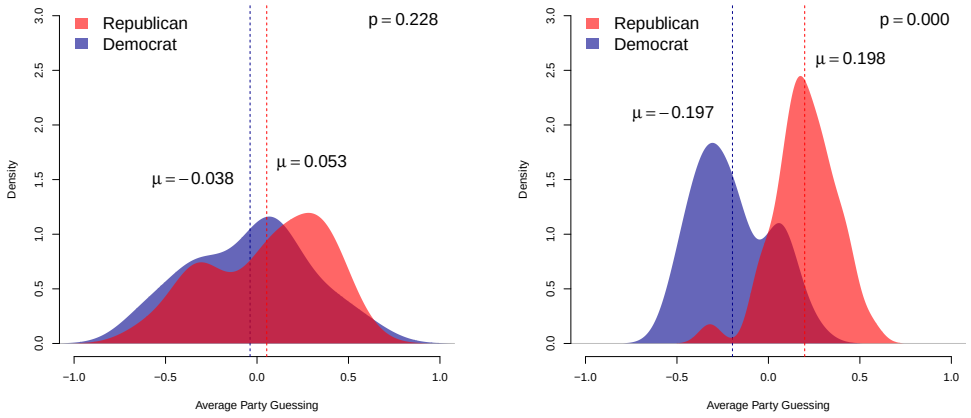
3.2 Dimensionality in Party Guessing

The resulting scores drawn from the party guessing experiments are presented in Figure 2. The figure displays densities of the guessing scores for positive ads in Figure 2(a) and negative ads in Figure 2(b) for the CCES sample, and similarly in Figure 2(c) and Figure 2(d) for the MTurk study. (In all figures, ads aired by Democrats are in blue, while those aired by Republicans are red.) The x-axis in these densities indicates the average partisan guessing score for the ads, with -1 representing all Democratic guesses, and +1 all Republican guesses. Negative attack ads are 'flipped' for presentational clarity, so that negative values indicate more guesses that the target is a Republican rather than a Democrat (hence a Democratic attack ad), while positive values indicate more

[12]The choice of the name Clark was to insure a common baseline that would be constant across voter inferences. There is evidence that voters can and do infer party from the gender or race of candidates (e.g., Goggin et al. 2015). This partisan effect may be minimized in using a generic white male candidate name. To the degree the name Clark biases things in the Republican direction, the bias is nevertheless constant across all the guessing experiments. Future experiments will randomize the name and gender of the candidate to insure that potential interactions are not a concern.

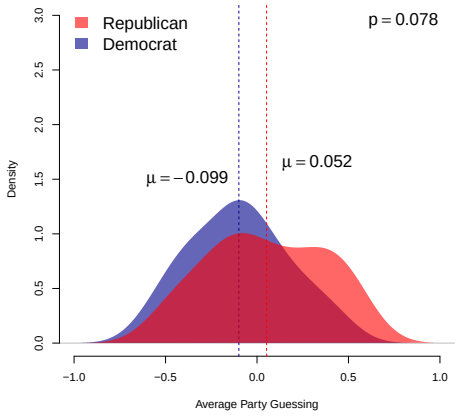
guesses that the target is a Democrat (hence a Republican attack ad). There is clear and meaningful variation across the scales, indicating that respondents are not just guessing randomly. At least on face, respondents perceive some of the ads to be strongly associated with each of the parties. Yet, quite a few of ads, and especially positive ones, are not able to be classified, or correctly classified to a party.

Figure 2: Density Plots of Guessing Scores for Positive and Negative Ads in the CCES and MTurk Samples

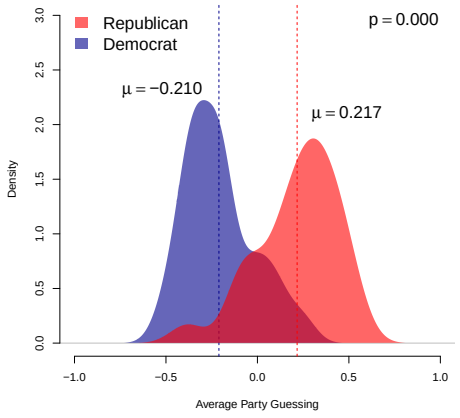


(a) Positive – CCES

(b) Negative – CCES



(c) Positive – MTurk



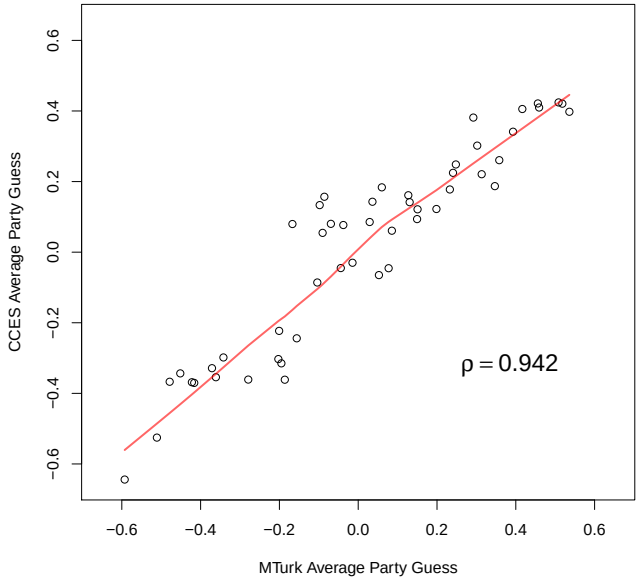
(d) Negative – MTurk

It is immediately obvious from these densities that positive and negative ads elicit very different patterns of guessing. Subjects are much better able to sort negative ads into party groupings, and to do so accurately. In comparison, positive Democratic and Republican ads are indistinguishable on average in the minds of survey respondents. In other words, there are positive Republican ads that appear to signal very consistent Democratic messages, and Democratic ads that signal consistent Republican messages. Thus, partisan signals in positive ads are much less "clear" (or consistent) than these are in negative ads. Henderson (2015) offers a theoretical explanation for this finding. Accordingly, positive ads aim to present candidates as relative moderates, by appealing to issues that are counter-stereotypical to party. This effort is meant to combat negative attacks, which seek to clarify opponents as consistent or extreme partisans.[13] Most relevant from a measurement perspective, (some) text-based scalings of ideological speech could have a difficult time recovering this particular dimension. These methods are better equipped to discern differences rather than similarities in speech across parties, and might struggle in contexts, like campaigns, where partisan candidates talk like each other.

Another notable feature is that the guessing scores look very similar in the MTurk and CCES samples. This similarity can be seen more clearly in Figure 3, which presents a scatterplot of average party guessing for the 50 overlapping ads included in both the CCES and MTurk frames. As seen, the distributions of guesses in both frames are highly correlated at $\rho = 0.94$, and statistically indistinguishable. Notably, this scatterplot presents unconditional averages *without adjusting for any differences between the samples on covariates*. This latter invariance is quite remarkable. There is no *a prior* expectation that these distributions be so similar. If anything, the substantial differences between subjects recruited into both surveys should yield important deviations in the guessing scores. Table 1 shows that subjects in the CCES and MTurk samples differ on virtually

[13]This distancing effect is not driven by efforts to avoid issues. Indeed, positive ads in these experiments contain equal amounts of issue information as do negative ads. Preliminary evidence suggests this effect is driven mostly by issue selection strategies.

Figure 3: Scatterplot of Party Guessing Scores On Overlapping Ads Between the CCES and MTurk Samples



every covariate collected in both surveys. MTurk respondents are 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. The last two differences (Democratic PID and liberalism) seem likely to skew party guessing in ways that could seriously bias comparisons across the two sample frames. It is possible these particular differences just happen to cancel out in the aggregate, but that other differences would be problematic. Additional replications can clarify this. Yet, this invariance, alternatively, may stem from the nature of the guessing task itself in tapping information that leads very different voters to make the same judgements in the aggregate.[14]

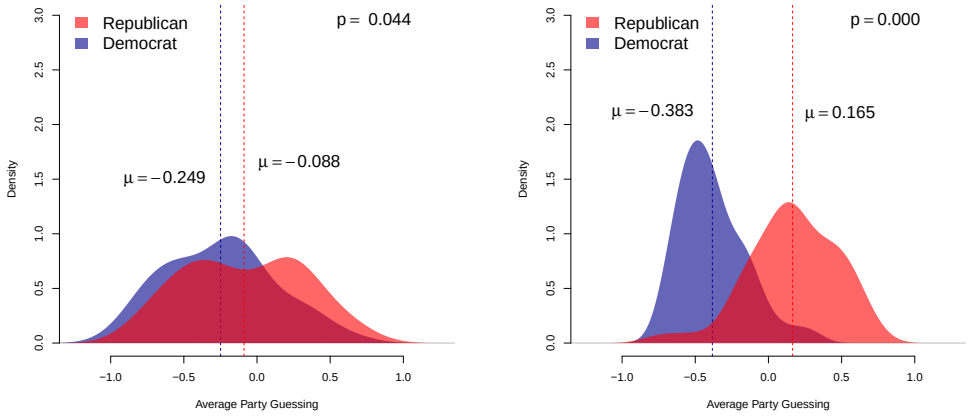
[14]Some differences do emerge between survey frames. MTurk subjects are less likely to answer 'Not sure', which increases the variance of the scores. There is also a slight shift towards Democratic guessing for positive ads and Republican guessing for negative ads, due to having nearly twice as many Democratic than Republican identifiers in the MTurk sample. Importantly, these shifts do not interact with any features of the ads, so that the rank-order of guesses is essentially invariant.

Table 1: Descriptive Statistics of Covariates Across CCES and MTurk Samples

	CCES_μ	MTurk_μ	p-value
Age	50.031	35.260	0.000
Female	0.531	0.474	0.000
White	0.734	0.769	0.001
Black	0.124	0.066	0.000
Hispanic	0.074	0.054	0.000
Asian	0.025	0.068	0.000
Other Race	0.043	0.037	0.238
Registered	0.896	0.906	0.161
Turnout	0.774	0.734	0.000
Education	2.638	3.154	0.000
Income	6.203	5.438	0.000
News Interest	3.069	1.023	0.000
Know House Majority	0.679	0.757	0.000
Correct Party Placement	0.793	0.918	0.000
Self Placement	0.082	-0.390	0.000
Party Identification	-0.400	-0.810	0.000
Presidential Vote	-0.100	-0.343	0.000

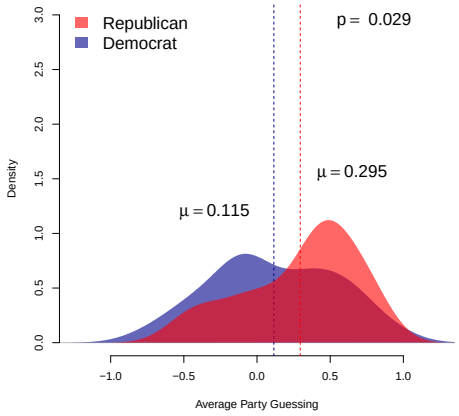
To bolster the latter interpretation, I stratify party guesses by PID. This stratification can illuminate whether and how partisans guess differently from each other, and especially for in- versus out-party ads. Such differences could be a serious problem if partisans evaluate ads differently depending on what features appear, as is commonly the case with party identification bias (e.g., Malhotra and Kuo 2008; Zaller 1992). Yet, if the bias in partisan guessing is simply additive, and partisans agree on the rank-ordering of ads, then this would be a minor issue in scaling. Figure 4, presents these guessing densities stratifying on PID. Responses are combined for both MTurk and CCES samples. In the positive frame, shown in Figure 4(a) and Figure 4(c), we see Democrats and Republicans are both more likely to believe that the ads are featuring candidates from their own party. Both distributions, however, shift together. Also, party identifiers are slightly better than random guessing at discerning positive ads aired by each party's candidates. Yet, the

Figure 4: Density Plots of Guessing Scores for Positive and Negative Ads, By Party Identity

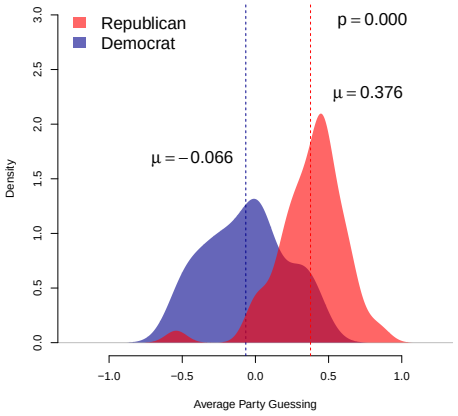


(a) Positive – Democratic PID

(b) Negative – Democratic PID



(c) Positive – Republican PID



(d) Negative – Republican PID

difference in these average differences between partisan ads as scored by Democratic and Republican identifiers is very small, and statistically null ($p = 0.85$). The shapes of the distributions are also similar, and highly correlated ($\rho = 0.84$) with each other.

In the negative frame, as seen in Figure 4(b) and Figure 4(d), partisans are more likely to guess that their *party opponents* are under attack, following something like a valence of logic of negativity (Goggin et al. 2015). This results in a shift in guessing

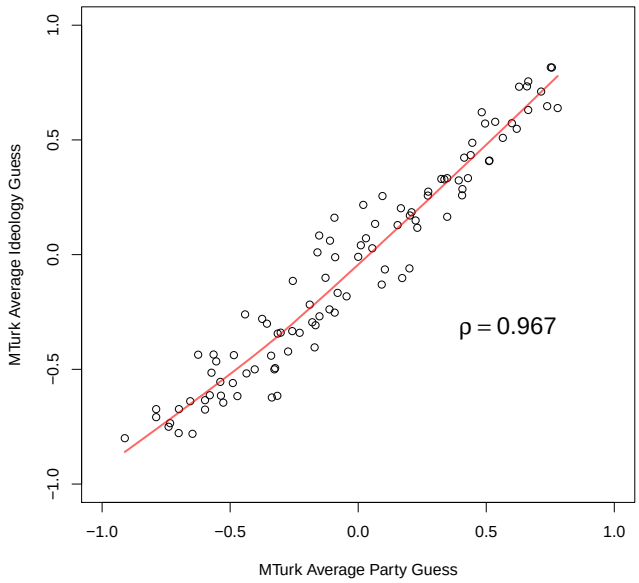
very similar to that observed for positive ads. Not surprisingly, partisans are also able to discern attacks aired by Democrats from those by Republicans. These scores again are highly correlated ($\rho = 0.87$) with each other across PID categories. And similar to the positive frame, the difference-in-differences between the way partisans guess Democratic and Republican negative ads is nominal and statistically null ($p = 0.134$).[15] Overall, in spite of these additive partisan shifts, the rank ordering of negative *and* positive guesses is statistically identical for Democratic and Republican identifiers

This fundamental invariance, even when comparing guesses across party identifiers, is an important feature of this scaling method. This finding points to the internal validity of the inferential task. 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 image 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.

Party guessing is stable across survey frames, and produces consistent scores even amongst different subsets of respondents, like partisans. Yet, how exactly do we interpret these average guessing scores, especially in ideological terms? I address this question using ideological rather than partisan labels in the guessing frame *Frame F* as described above. Recall that the task involves randomly assigned respondents classifying ads based on if they think the statements feature a 'Liberal' or 'Conservative' candidate, rather than a 'Democrat' or 'Republican'. A scatterplot of the resulting guesses are presented in Figure 5. The plot presents average party guesses from the MTurk *Frame C* on the

[15]The shapes of these distributions are also very similar according to a Kolmogorov-Smirnov (ks) test, at $p = 0.87$ for positive ad differences, and $p = 0.18$ for negative ad differences.

Figure 5: Scatterplot of Scores Guessing Party and Ideology



x-axis, and average ideology guesses in *Frame F* on the y-axis. As seen, the correlation between the two scores is $\rho = 0.97$, and the distributions are indistinguishable from each other. This finding indicates that the likelihood that an ad is Democratic is identical to the likelihood the ad is liberal in the minds of voters. In this way, party guessing captures meaningful ideological information in ads that voters can identify.

One challenge in interpreting this finding is that it could result from voters inferring party in the ad, and then attaching some ideological label to it, simply from knowing that Democrats are liberal and Republicans are conservative. (The opposite inferential direction is also possible.) This would limit the interpretation of these scores as capturing a partisan instead of an ideological dimension. An implication of this sort of 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. Ultimately, we cannot peer into the minds of voters to see how they perceive party and ideology

in ads. But this evidence demonstrates that ideological and partisan guessing produce indistinguishable measures of ad content.[16]

3.3 Validating the Ideal Point Measures

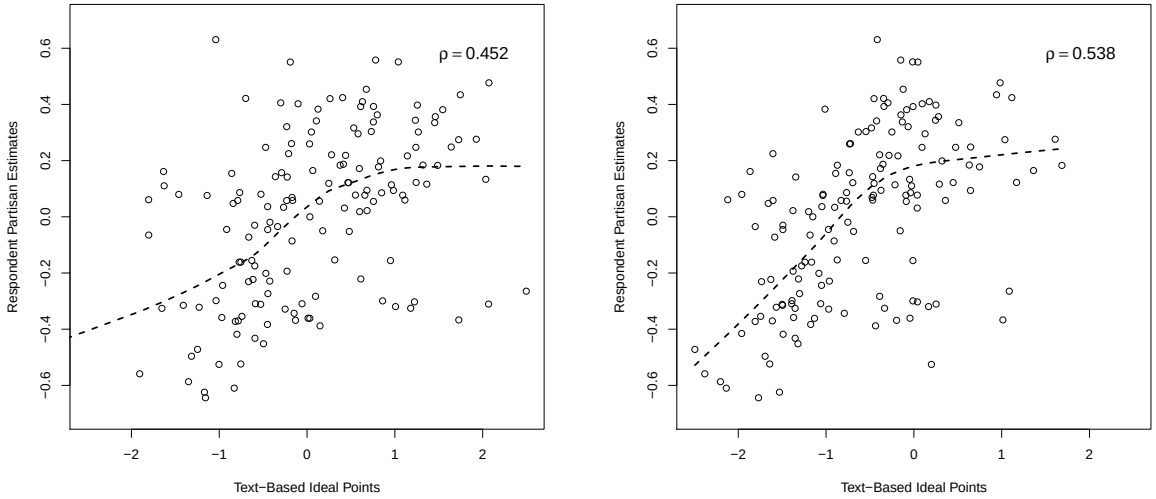
These ad guessing scores appear to be invariant to whether subjects are asked to guess the party or ideology of candidates, or differ along a range of plausibly important covariates, like PID, education or political interest. These findings strongly suggest that, through an inference task, very different voters at least in the aggregate, can agree on the ordering of ads based on their partisanship, which is indistinguishable from its ideological content. This adds much credence to this method of scaling built upon the perceptions voters have about ads. Though consistent across survey contexts, it is possible that these perceptions are not the best way to scale the information in ads. Voters could be collectively myopic, consistently misjudging the content of ads. Or voters may behave in surprising ways following exposure to ads given how these are scaled. Hence it is important to validate this method of scaling through a number of tests, with the particular goal of comparing this approach against alternatives, in this case *Wordscores* and *Wordfish*.

I first compare party guessing to analogous scales produced using just the ad text through *Wordscores* and *Wordfish* as described above (and outlined in the Appendix in more detail). I present scatterplots of these comparisons in Figure 6. As shown, party guessing correlates positively with *Wordscores* at $\rho = 0.45$ in Figure 6(a), and similarly with *Wordfish* at $\rho = 0.54$ in Figure 6(b). Though these correlate, there is considerable non-linear variation here, and especially as these scales approach their midpoints at zero.[17] Hence text-based approaches produce related, but very different scales compared with party guessing. One likely explanation explored in Henderson (2015) is that text-based

[16]This is also true with ideal point measures of legislative behavior which are consistent with either interpretation that these capture consistent party behavior or ideological extremity.

[17]Non-linearity is found for both positive and negative ads, though negative ads are more linear.

Figure 6: Scatterplots of Party Guessing Measures Compared to Both *Wordscores* and *Wordfish* Alternatives



(a) Comparing *Wordscores*

(b) Comparing *Wordfish*

approaches may have a difficult time appropriately scaling ads when these are aimed at strategically presenting counter-stereotypical partisan information (i.e., candidates strive to present themselves as moderates). 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 alternatives is to identify whether text- or human-based approaches perform better in explaining the way ads are targeted in congressional districts, or the effects ad exposure may have on voter behavior or attitudes. For the latter, 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 7. The candidates evaluated in the experiment are

Figure 7: Experiment Protocol for Candidate Vignettes

In a recent U.S. House election, **Tom Hill** and **Mark Meadows** ran on the following positions on the budget and taxes:

	Middle Class Tax Cut Act. Extend tax-cuts only for individuals *with incomes below $200,000*.	**Simpson-Bowles Budget.** Cuts *Medicare and Defense* spending to reduce the federal deficit.	**Tax Hike Prevention Act**. Extends tax cuts for all individuals *regardless of income*.	**Paul Ryan Budget**. Cuts *Medicare and Medicaid* to reduce the federal deficit.
Tom Hill	✓	✓	✗	✗
Mark Meadows	✗	✗	✓	✓

During the election campaign, **Tom Hill** also had this to say:

"We need to be energy independent. For a decade, I've been a leading voice in Congress for alternative energy solutions, biofuel research, geothermal energy, solar and next generation nuclear. In a time when Washington can't do anything, I got legislation signed into law to move energy efficient technologies out of the laboratory and into the marketplace."

If you could have voted in this election, which candidate would you have supported?

Mark Meadows
Tom Hill
Not sure

Mark Meadows and Tom Hill.[18] In the protocol, respondents are shown a brief preface, which includes the positions each candidate took on four roll call votes, along with

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

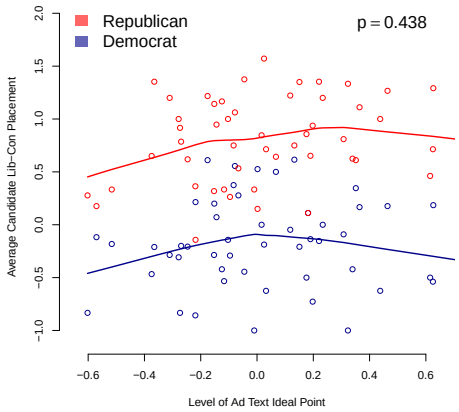
descriptions of the votes. The votes included were: Middle Class Tax Cut Act, Simpson-Bowles Budget, Tax Hike Prevention Act, and Paul Ryan Budget.[19] Below this policy information preface, respondents are then randomly shown an ad statement from one of the two candidates. The candidate who appears is randomly selected. The message that is then attributed to the selected candidate is also randomly chosen, drawn from the 50 positive ads scored previously in the CCES party guessing experiments.[20] 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.

Notably, 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. These positions should help inform voters about the overall ideological positions these candidates are likely to take on other issues, and may imply partisan information. However, it should be clear the latter is explicitly *not* being primed in the experiment. (Future experiments will explicitly prime party to see if ads can alter voter impressions even when partisanship is clarified.) The only direct policy information respondents receive is through these roll call vote positions, and the ad statements randomly attributed to candidates. Key to the design then, is that voters infer some policy information about the candidates through the roll call positions, which is then mediated by various ad statements that range from left to right as estimated by actual respondents through party guessing in another experimental sample. 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.

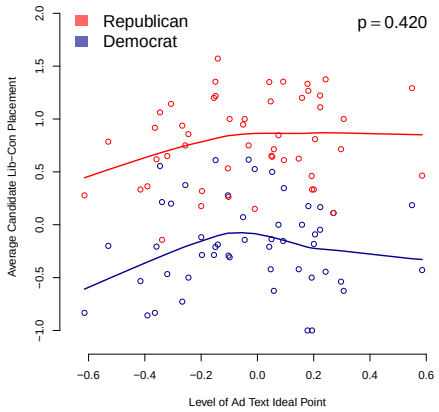
[19]These votes were chosen due to their salience in legislative and party politics, as well as their inclusion in the list of common content questions in the CCES. The latter allows a comparison between voter attitudes on these items and evaluations of the candidates later on. The issues were arrayed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

[20]An important cross-over element is used across CCES *Module A* and *Module B*, so that respondents never provide guesses for any ads that they could see in this candidate vignette experiment.

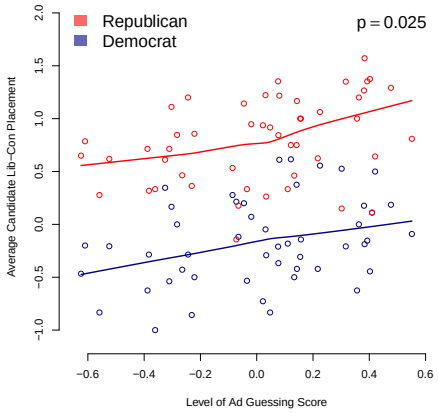
Figure 8: Influence of Ad Exposure on Voter Candidate Placement, By Method of Scaling Ads



(a) *Wordscores*



(b) *Wordfish*



(c) Party Guessing

The results of the vignette experiments are presented in Figure 8. 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

between the candidates on the ideological scale. The x-axis in these plots indicates the scaled location of each of the 50 ads, as scored by (a) *Wordscores*, (b) *Wordfish*, and (c) Party Guessing. 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 8(a) and Figure 8(b), neither of the text-based scales correlates 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 8(c), we see that the ideological information in ads, as scaled by other respondents, is a significant predictor of candidate placement. Ads that are rated as more liberal (conservative) through party guessing, when randomly aired by candidates, correspond with more liberal (conservative) candidate placements. Notably, this pattern holds for *both* the relatively liberal (Democrat) Tom Hill and the relatively conservative (Republican) Mark Meadows.[21] 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 some caution to the general use of these automated approaches to analyze the effects of messages received by voters.

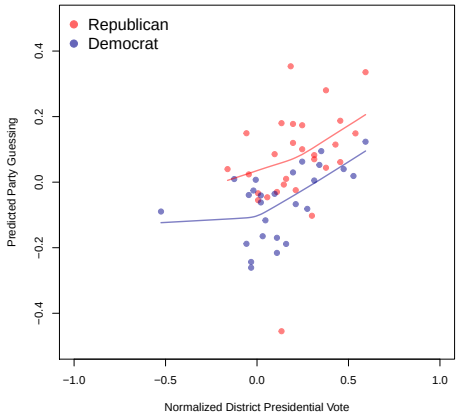
A final validity check of each scaling is to examine whether any of these correlate with the aggregate preferences of voters in the districts in which the ads were aired in 2008. To do this, I average the ad scalings for each candidate airing one of the 200 ads to the district level. These district ad positions are then compared to (normalized)

[21]Figure 8 just presents the bivariate correlations. These persist when adding both additional ad- and individual-level controls.

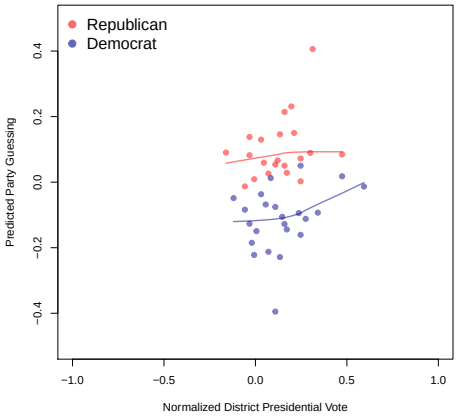
district presidential vote in the 2008 election. A scatterplot of this is presented in Figure 9. The x-axis in the figure is presidential vote, and the y-axis is average candidate-level positioning scaled using each method. As shown in Figure 9(a), there is a clear, positive association between conservative ad positions and Republican presidential choice, for both Democratic (blue) and Republican (red) House candidates as scaled by party guessing. Yet, this association is much weaker for *Wordfish* in Figure 9(c), and statistically zero for *Wordscores* in Figure 9(b). To the degree we expect candidates to campaign in ways that reflect district attitudes, this evidence suggests that party guessing measures best capture this targeting strategy.

A similar comparison is made here for negative ads. For party guessing in Figure 9(b), we see that negative ads polarize, but only weakly correlate with district vote choice. In comparison, both *Wordfish* and *Wordscores* scalings of negative ads strongly correlate with district vote choices as shown in Figure 9(d) and Figure 9(e). Yet, according to these measures, candidates' negative attacks appear somewhat unusual. In relatively liberal or centrist districts, both Republicans and Democrats attack their opponents as being too conservative, while in more conservative districts both parties attack their opponents as being too liberal. From a purely Downsian perspective this strategy could make sense (Downs 1957). Though in an era of polarized politics, it seems unlikely that conservative Republicans would attack Democrats as being too conservative, or liberal Democrats attacking Republicans as too liberal. Admittedly, the evidence here for negative ads is harder to interpret. But, under a standard view of the politics of negativity, it seems more plausible that these attacks would separate ideologically across parties, rather than converge at the district level (Geer 2006; Henderson 2015). Such a pattern is far more evident using party guessing scores, rather than the alternatives.

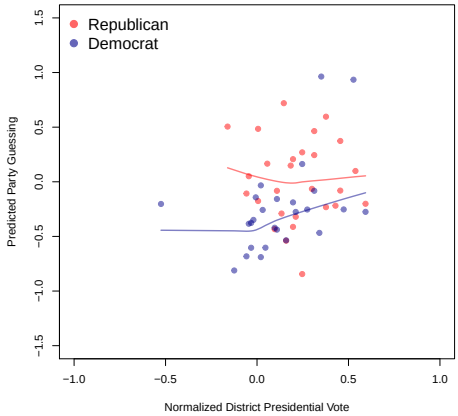
Figure 9: Scatterplot of District Presidential Vote and Candidate-Level Scores for Three Ad Scaling Methods



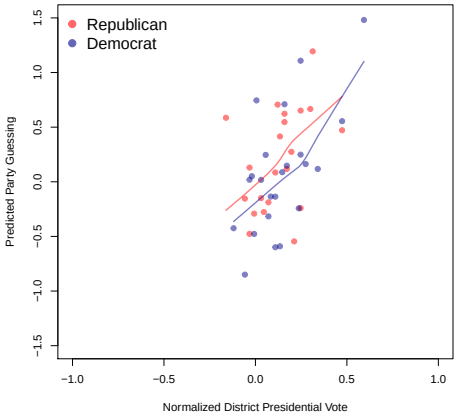
(a) Guessing - Positive



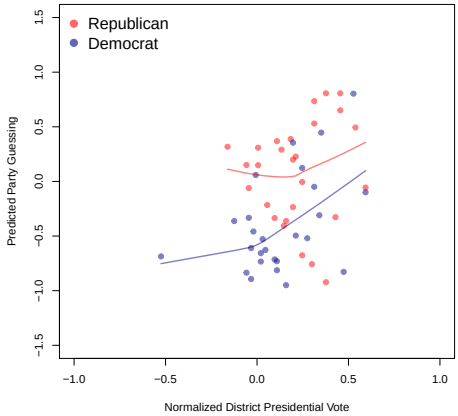
(b) Guessing - Negative



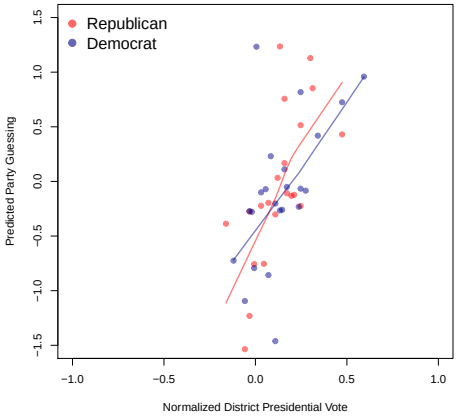
(c) Wordscores - Positive



(d) Wordscores - Negative



(e) Wordfish - Positive



(f) Wordfish - Negative

4 Supervised Learning to Predict Ad Guessing Scores

Party guessing scores appear to perform better in each of the above validation tests. These scores capture the information that voters seem to use when evaluating candidates airing the ad messages. These also best accord with the kinds of targeting strategies we might expect candidates typically employ in the campaign. Yet, constructing these scores can be somewhat costly since the approach requires survey respondents to offer many guesses for each ad.[22] Given that automated scaling only costs researcher time, for a large number of ads, it may be worthwhile to accept some measurement error from automation to obtain the benefits of efficiently scaling an entire corpus of ads. Alternatively, it might be possible to automate some part of the guessing task. Such an automation would involve guessing a portion of all ads, and then using supervised learning about those guesses to make predictions about the remaining ads. If feasible, this can reduce the scope of the guessing task, making it more efficient and less costly. In the ideal, this also could be accomplished by tuning the proportion of all ads required to be coded to obtain some pre-determined level of expected measurement error.

In this section, I explore whether party guessing scores can be used in such a supervised learning approach. In particular, I use the 'elastic net' to make predictions about guesses using covariance in ad words (Zou and Hastie 2005). I then develop simulations to evaluate how best to scale up the guessing task to make better predictions for an entire corpus just using a subset of ads. In addition to helping assess the above trade off between human-based and text-based approaches, these supervised simulations can help clarify whether the major challenge in scaling campaign advertising is due to data sparsity or strategic speech. If ad words can be used to successfully predict party guesses, this

[22] One helpful feature of guessing shown above is that the rank ordering of scores are apparently invariant to many individual-level factors that might be expected to influence guesses. A consequence of this is that guessing likely can be done using non-representative samples, such as those typically found in MTurk, to produce valid scores that are very similar to those recovered using high-quality, representative samples, such as the CCES. This reduces but obviously does not eliminate costs.

suggests that sparsity itself does not prevent automated approaches from recovering this particular partisan dimension, but rather that this dimension is distinct from how other automated approaches capture political disagreement in speech.

4.1 Using the Elastic Net to Predict Party Guessing Scores

Recall in the CCES study that 150 ads were originally scored using party guessing, while another 50 ads were left out. These latter 50 ads were later scored in the MTurk study. By separating this guessing task, it is possible to make supervised predictions about these left out ads using the CCES scores *before their outcomes were collected* through MTurk. Since the MTurk survey was not put in the field until after predictions were recovered for the CCES ads, this provides a very meaningful test of the approach.

For an automated learner I utilize the elastic net, which is a powerful way to make supervised predictions about guessing scores $S^{(d)}$ given the words W_d in d documents (Zou and Hastie 2005). The learner regresses $S^{(d)}$ on W_d using a standard linear model:

$$S^{(d)} = \gamma' W_d + \epsilon_d$$

As with most text data, the number of words typically far exceeds the number of documents, so that this model is not identified. Manually pre-selecting words so the model is just-identified is an option, but typically produces unreliable predictions. By placing constraints on the regression coefficients, the elastic net can identify the above model (Zou and Hastie 2005).

The elastic net combines both the lasso and ridge regression. This simultaneously constrains all regression coefficients γ_j so these do not grow astronomically, while at least some are selected to be non-zero.[23] Ridge regression regularizes coefficients by minimizing

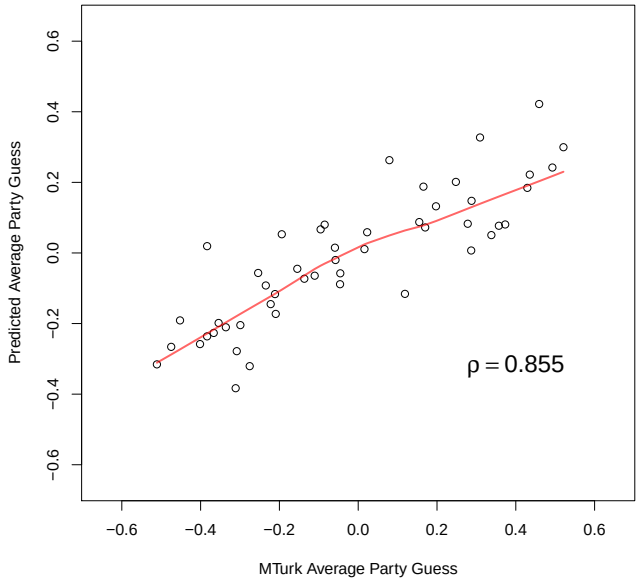
[23] In this sense, the ridge and lasso are like frequentist alternatives to using priors in Bayesian analysis. The elastic net is often preferred over the lasso since it has the general feature of stabilizing lasso

the sum of least squares, $\sum(S^{(d)} - \gamma' W)^2$ subject to the constraint $\sum \gamma_j^2 \leq t_{L2}$, while the lasso imposes the constraint $\sum |\gamma_j| \leq t_{L1}$. Elastic net regularization then combines these together using the following constrained optimization

$$\hat{\gamma} \quad = \quad \arg\min_{\gamma} \sum \left(S^{(d)} - \gamma' W\right)^2 + a\lambda \sum \gamma^2 + (1-a)\,\lambda \sum |\gamma|\,,$$

where a common λ constraint is used, and its influence is apportioned by a. The payoff is that the elastic net can be estimated even with highly sparse data, since the constrained optimization allows the identification of γ even when the word matrices are not full rank.

Figure 10: Scatterplot of MTurk Party Guesses Compared to Elastic Net Supervised Predictions for 50 Testing Set Ads



For a given level of a and λ, the coefficients $\hat{\gamma}^d_{a,\lambda}$ can be used to make predictions about documents d' left out of the prior estimation stage. Yet, there are now an infinite number of regression solutions. Thus, a and λ parameters must be tuned within some bound to

regularization, which can be haphazard due to the absolute loss constraint.

minimize prediction error. This is usually done by exploring a range of parameters and then using *cross-validation* for each combination of a and λ. Cross-validation evaluates the prediction accuracy using a subset of training documents, to pick the levels of a and λ, that minimize error on $S^{(d)}$ on all the training documents. Here I use 20-fold cross-validation, dividing the 150 documents into $K = 20$ groupings of $N \approx 7.5$ ads. For the kth grouping, at fixed $\{a, \lambda\}$, I estimate word parameters $\hat{\gamma}^{k'}_{a,\lambda}$, using all the ads in the other 19 groupings denoted by k', where $k \cup k' = d$, and $k \cap k' = \emptyset$. A predicted value for $\hat{S}^{(k)}$ is then estimated as $\hat{\gamma}^{k'}_{a,\lambda} \times W_k$, which is repeated for each k subset. A *mean square error* (MSE) summary statistic is produced for each level of $\{a, \lambda\}$, which evaluates the expected accuracy of the predictions on the k subsets modeling words in the other k' ads:

$$\begin{aligned} \text{MSE}(a, \lambda) &= \frac{1}{KN} \sum_k \sum_{i \in k} \left(S_i^{(k)} - \hat{S}_i^{(k)} \right)^2 \\ &= \frac{1}{KN} \sum_k \sum_{i \in k} \left(S_i^{(k)} - \hat{\gamma}^{(k')}_{a,\lambda} W_i \right)^2 \end{aligned}$$

The level of a and λ are chosen that minimize $\text{MSE}(a, \lambda)$ over the full range of parameters explored. The range of parameters explored here are $a = \{0, 1\}$ and $\lambda = \{0.01, 10\}$, and the optimal values are $a = 0.3$ and $\lambda = 0.039$.[24]

Once the optimal word coefficients are identified in the training set, these coefficients are then used to make predictions for the testing set $\hat{S}^{(d')}$. Figure 10 presents a scatterplot of these predictions for the 50 held out ads compared to their guessing scores $S^{(d')}$ recovered in the MTurk sample. (Again the MTurk scores are unconditional on respondent covariates.) As seen, the correlation between actual and predicted guesses is high at

[24]When $a = 1$ the above regression optimization produces ridge regression coefficients, while $a = 0$ produces lasso coefficients. As $\lambda \to 0$, both the lasso and ridge regression coefficients approach a 'standard' least squares optimization. As $\lambda \to \infty$, ridge regression sets all γ coefficients (except the intercept) to zero, while the lasso performs feature selection, setting a subset of γ to 0. As λ increases under the lasso, the non-zero coefficients tend to increase in a non-smooth fashion, which can degrade predictive accuracy, particularly if an insufficient range (or granularity) of λ is explored. One benefit to using the elastic net implementation is that the ridge square loss constraint tends to smooth out the growth of lasso coefficients as λ increases. This often produces more accurate predictions.

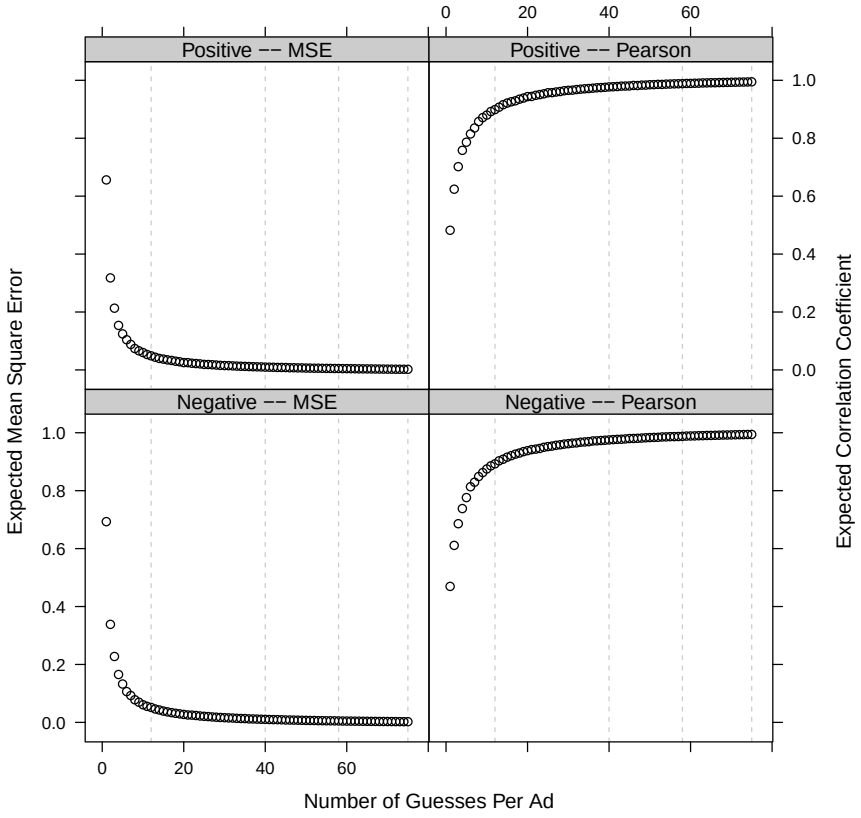
$\rho = 0.86$. These distributions are statistically indistinguishable (e.g., using both a t-test and ks-test). The supervised learning approach, thus performs quite well in predicting partisan guesses from the covariance of words in ads. In one sense this should not be too surprising since those are the same words that respondents use to make judgements about the partisanship of ads. Yet, this also suggests that the sparsity in ads data is not so limiting as to prevent *any* machine learning method from appropriately scoring ads on an ideological scale. In other words, the other automated scaling approaches may provide consistent estimates of some political dimension, albeit one distinct from how voters perceive the ads.

4.2 Simulations to Explore the Predictive Properties of Ad Guessing

The supervised approach implemented above yields predictions that are highly correlated with actual guessing. Yet, this success may depend on the proportion of ads used in the training set relative to the rest of the corpus. Indeed in the above case, 150 ads are used to make predictions about 50 others. In order to scale up this predictive task to many more ads, it will likely be necessary to also scale up the number of ads included in the training set, and thus to be scored through surveys. One important question here is how many guesses per ad is needed to produce high quality scores. A second question is how many ads need to be included in the training set to produce high quality predictions on the remaining ads. The goal in scaling up this method is to simultaneously minimize the number of guesses per ad and the number of ads guessed, while maximizing the quality of the scores. To address this goal, I implement two simulations that can help answer both of the above questions.

In both the CCES and MTurk surveys, a large sample size was used to insure that there were at least 100 guesses per ad in expectation. Yet, it is possible that fewer guesses per ad would do reasonably well in measuring ad partisanship. If so, then a

Figure 11: Simulation I: Convergence in Guessing Scores as the Number of Guesses Per Ad Increases



way to gain efficiency would be to minimize the number of guesses needed to yield a certain level of accuracy in guessing. A way to identify this ideal number of guesses is to simulate the error in the scores that would be produced if $m < 100$ guesses were utilized. For this simulation, m number of guesses are randomly selected from the actual distribution of survey responses for the 200 ads, for each level of m going from 1 to 75. An average score is produced for each ad using m number of randomly selected guesses $\tilde{S}_m^{(d)}$. This score is then compared to the actual scoring using all the guessing data $S^{(d)}$. Two summary statistics are taken from this comparison: the mean square error and the correlation between each mth score and the full score. This simulation is repeated 1,000

times through bootstrap sampling. For each s bootstrap sample, m number of guesses are randomly chosen for each m level, to produce a score $\tilde{S}^d_{s,m}$. The average mth score is taken over these s samples to produce $E[\tilde{S}^{(d)}_m] = \frac{1}{1000}\sum_s \tilde{S}^{(d)}_{m,s}$. The correlation and error summaries are then computed using this quantity.[25]

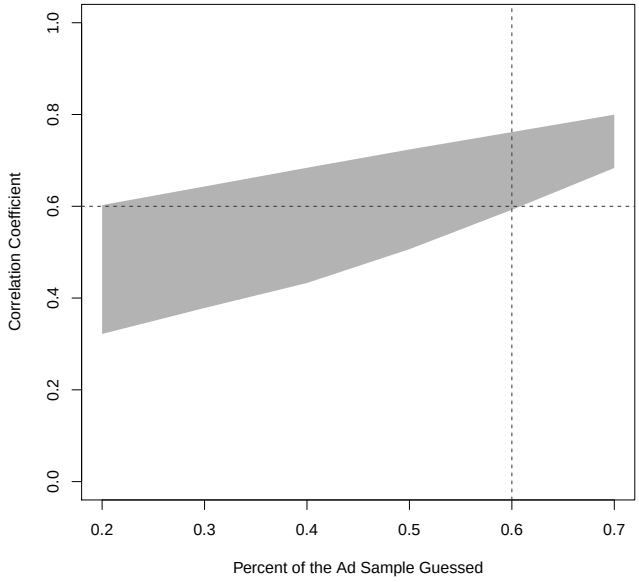
The results of the simulation are presented in Figure 11. The main finding is an obvious one: as the number of guesses increases, the resulting scores converge on those recovered using 100 expected guesses. Yet, these simulations can also identify the level of m needed to obtain some desired amount of convergence. For example, to recover a MSE of no more than 0.01, at least 40 average guesses are needed. This would be expected to yield a correlation of $\rho = 0.98$. A stricter standard would require a MSE of 0.005. This would require 58 guesses and would produce an expected correlation of $\rho = 0.99$.[26] Hence it turns out to be unnecessarily costly to require more than 60 guesses per ad, and even 40 guesses per ad seems sufficient to produce highly reliable scorings.

Another way to gain efficiency is to identify the smallest proportion of total ads necessary to be guessed that will result in sufficiently accurate supervised predictions. This question is a bit trickier than the above. It is impossible to know what proportion of 1,662 ads is needed to produce high quality predictions for a left-out set of these total ads without actually scaling all them. One way to approximate this is to consider the 200 ads above as the whole corpus of ads. Then random subsets of various sizes (e.g., 100, 125, 150, 175) of these 200 ads can be used to make predictions about the remaining left out ads. This is conducted in a second simulation where 1,000 samples s are drawn for each discrete proportion, $p = 0.2, 0.3, ..., 0.7$, of the 200 total ads. The $p \times 200$ ads are used to make predictions about the $(1-p) \times 200$. These predictions are made using

[25]Perhaps a simpler way to describe this bootstrap simulation is an effort to explore the asymptotic convergence of average guessing as the number of guesses is allowed to increase. Each bootstrap is analogous to repeating the experiment allowing for m expected guesses in the design. Bootstrapping smooths out estimates of the asymptotic properties of convergence.

[26]If a standard MTurk survey costs around 0.07 per guess, then fielding 40 instead of 100 guesses per ad, to score 1,662 ads, would cost around $4,643 and save an additional $6,980.

Figure 12: Simulation II: Proportion of Total Ads Guessed to Elicit a Pre-Determined Correlation Level in Prediction



the elastic net approach described above to minimize the MSE on the training set, here the random sample of p ads. The correlation for each s sample prediction and the actual guessing score is computed, along with the bootstrap 95% confidence interval of these.

Figure 12 displays the results of this simulation. The x-axis indicates the percent of the total number of ads used in the supervised learner, and the y-axis indicates the resulting correlation for each bootstrap sample and level in m, comparing supervised predictions to actual guessing scores. (The shaded grey area indicates the 0.05 and 0.95 confidence interval around the expected correlation for the bootstrap sample predictions.) The main finding from the simulation is that a conservative bound (i.e., the 0.05 lower bound) is roughly linear in the percentage of ads to be guessed. In other words, in expectation, at least 60% of the 200 ads need guessing outcomes from surveys to produce a correlation of $\rho = 0.6$ in comparing the predictions on the other 40% of left out ads,

and the actual guessing scores.

This is a conservative bound in being at the low end of the confidence interval. It is certainly possible to do better than $\rho = 0.6$ using 60% of the data, and in fact such is the case 95% of the time. Using this conservative benchmark, however, insures against the possibility that some idiosyncratic features of the included ads induce significant error in prediction that would not otherwise be observed. The simulation also cannot speak to the accuracy in predictions as the total size of the corpus increases. It is possible that as the total number of ads increases, the sparsity in (the most predictive subset of) the data declines, enhancing the accuracy in prediction. Yet, the opposite could easily be the case. Thus, using this conservative bound also can hedge against the concern that added sparsity will reduce the quality of prediction as the size of the corpus increases. Overall, these simulations demonstrate that it is quite feasible to use of a combination of automated and hand-coding analysis to produce a highly valid, reliable and powerful tool of ideological measurement applied here to the scaling of ads.

5 Conclusion

In this paper, I develop an experimental approach to scale the ideological content of political ads. I randomly assign ads to survey subjects who are asked to guess the party and ideology of featured candidates. Ads are then scaled as their expected partisan guessing score. I find that this partisan score is analogous to an ideological dimension, and can be accurately predicted in a supervised learning framework using the words in ads. I then show that these scores out-perform standard automated approaches to scaling text through a number of validations. Finally, I demonstrate that this party guessing approach can be scaled up to a code a large number of ads in a relatively cost-effective manner, that is likely to produce less measurement error than the alternatives. This method represents an important advance in the scaling of text. This is particular so in contexts laden with

strategic speech that often frustrates standard models of ideal point estimation.

Throughout the paper, I argue that this ad guessing method is likely to be most effective when speech is strategic or sparse in ways that limit the accuracy of automated approaches. Even when automated scaling works well, guessing can be a useful mode of validation to insure that word-based ideal points have some grounding in realism and are meaningful to voters. This latter point is especially important when working through the logic of validating ideal point scores. Doing so fundamentally requires making a theoretical or substantive assumption about the relationship between the objects being scaled, and the objects used to validate the new scaling, i.e., the ideal point tautology. The logic assumes one scaling is 'correct' in a deep sense, *and* is the appropriate benchmark to evaluate other scales. Under this logic, any deviation between a new scale and this benchmark indicates poor measurement, as opposed to a poor assumption about the likely distribution of the new scale, conditional on the benchmark. Party guessing, as a mode of validation, can clarify how strong this assumption is in a number of contexts.

For example, there is heated disagreement about how to model position-taking in campaign competition. Using legislative voting to predict candidate advertising makes the strong assumption that ads are meant to reflect prior legislative positions. An alternative is that politicians pander to voters or converge on median voter preferences in the campaign. Each of these assumptions suggests very different ways of validating ad ideal point scores. Any (or all) of these could be incorrect, *and* the only way to evaluate these assumptions empirically is by comparing an otherwise validated scaling of ads to scaled legislative positions, or voter and donor attitudes. Hence, traditional modes continue in an endless circularity between measurement and theory.

Party guessing provides a clear path out of this tautology. Guessing simply measures whatever partisan information voters actually observe in ads. Hence this can provide an independent basis (e.g., replicability, exploring bias in the scale from respondent char-

acteristics) to judge the quality of the resulting scaling. If campaign positioning reflects an appeal to centrist voter preferences when these positions are 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. To a significant degree, party guessing scores can clarify important theoretical questions, without needing to reference other scalings to support its validity. Thus, in breaking the tautology, guessing can enhance the information value of comparing automated scalings to each other by providing evidence about the assumptions guiding those comparisons.

Beyond an ideological scaling of ads, the inferential approach outlined here can be generalized to measure a much wider array of dimensions contained in speech and text data. While the application of guessing in this paper is to scaling ads in the U.S. two-party context, in principle guessing can be expanded to other kinds of speech or text, and in other partisan or political contexts. An important feature of the scaling approach is that survey respondents, for at least some tasks, can be shown to produce highly reliable and internally valid scores using crowdsourcing services that make recruiting a large number of subjects very affordable. The limiting factors in using this approach are likely to be the complexity of the survey instruments, and the cost of using humans to code text. When both factors are low, it may make sense to use crowdsourcing 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 immensely 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 far more powerful and effective tool of measurement, that can compliment and improve the analysis of text data.

References

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

Beauchamp, Nick. 2010. "Text-Based Scaling of Legislatures: A Comparison of Methods with Applications to the US Senate and UK House of Commons." Working Paper.

Benoit, Kenneth, Michael Laver, Drew Conway and Slava Mikhaylov. 2012. "Crowd-Sourced Data Coding for the Social Sciences: Massive Non-Expert Coding of Political Texts." Presented at New Directions in Analyzing Text as Data Conference.

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.

Budak, Ceren, Sharad Goel and Justin M. Rao. 2015. "Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis." Working Paper.

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

Diermeier, Daniel, Jean-Francois Godbout, Bei Yu and Stefan Kaufmann. 2011. "Language and Ideology in Congress." *British Journal of Political Science* 42(1):31–55.

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.

Geer, John G. 2006. *In Defense of Negativity: Attack Ads in Presidential Campaigns*. Chicago: University of Chicago Press.

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.

Groseclose, Tim, Steven D. Levitt and James M. Snyder, Jr. 1999. "Comparing Interest Group Scores Across Time and Chamber: Adjusted ADA Scores for the U.S. Congress." *American Political Science Review* 93(1):33–50.

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." Working Paper.

Hopkins, Daniel J. and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1):229–247.

Kim, In Song, John Londregan and Marc Ratkovic. 2014. "Voting, Speechmaking, and the Dimensions of Conflict in the US Senate." Presented at the Annual Meeting of the Midwest Political Science Association.

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." Presented at New Methodologies and Their Applications in Comparative Politics and International Relations Conference.

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.

Malhotra, Neil and Alexander Kuo. 2008. "Attributing Blame: The Public Response to Hurricane Katrina." *Journal of Politics* 70(1):120–135.

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.

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.

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.

Spirling, Arthur. 2012. "U.S. Treaty Making with American Indians: Institutional Change and Relative Power, 17841911." *American Journal of Political Science* 56(1):84–97.

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

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

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

Zou, Hui and Trevor Hastie. 2005. "Regularization and Variable Selection Via the Elastic Net." *Journal of the Royal Statistical Society, Series B* 67(2):301–320.

A Appendix

A.1 More Details on Automated Approaches to Scaling Text

The *Wordfish* model of speech takes the following form:

$$\begin{aligned} W_{ij} &\sim Poisson(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\alpha_i + \psi_j + \beta_j \times \omega_i). \end{aligned}$$

The λ_{ij} term is a Poisson parameter increasing in the number of times legislator i utters word j, measured by W_{ij}. The term α_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, that is, the degree to which the word is likely to be used mostly by liberals rather than conservatives, or the reverse. This discrimination parameter plays an important role in tuning how much influence the legislator ideal point ω_i plays in determining the frequency that i speaks word j. A way to represent this model is in terms of a word 'cutpoint', $m_j = -\frac{\psi_j}{\beta_j}$, which defines the point in space where a legislator would be indifferent between choosing to utter a word and remaining silent.

An alternative to this unsupervised utility approach is supervised learning. One variant of this method is to build a dictionary of 'liberal' and 'conservative' words, and then score documents based on the frequency in which these ideological words are used (Beauchamp 2010; Laver et al. 2003; Lowe 2008). This approach is analogous to predetermining the above discrimination parameters β_j for words (say at -1 and 1), and then estimating ω_i in the above model, treating the j terms as fixed data (and ignoring ψ_j). A common implementation of this dictionary approach works at the level of scoring words, hence taking the name *Wordscores* (Laver et al. 2003). Reference texts C and L are first chosen by the researcher to represent canonical conservative and liberal statements. The target is to estimate $p(C|W_j)$ and $p(L|W_j)$, or the probability a document is

conservative or liberal given the use of W_j in these pre-determined statements, and then to summarize over all j for each document. Denote $W_j^{(C)}$ to be the count of W_j appearing in document C, and analogously for $W_j^{(L)}$. Word proportions then can be constructed (following an unnormalized formulation used in Beauchamp (2010)) as:

$$p(C|W_j) \quad = \quad \frac{W_j^{(C)}}{W_j^{(C)} + W_j^{(L)}}.$$

These proportions capture how likely it is that word j appears in a conservative text. A word score S_j is then built, assuming a weight and polarity for discrimination in W_j, commonly designated ± 1. This score is

$$S_j \quad = \quad p(C|W_j) - p(L|W_j).$$

These individual word scores can then be used to summarize an entire document D as:

$$S^{(D)} \quad = \quad \sum_j \frac{W_j^{(D)}}{W^{(D)}} \times S_j.$$

Here $W_j^{(D)}$ is the count of word j in D, and $W^{(D)}$ is the total number of words in D.

A.2 Differences in Positive and Negative Guessing Frames

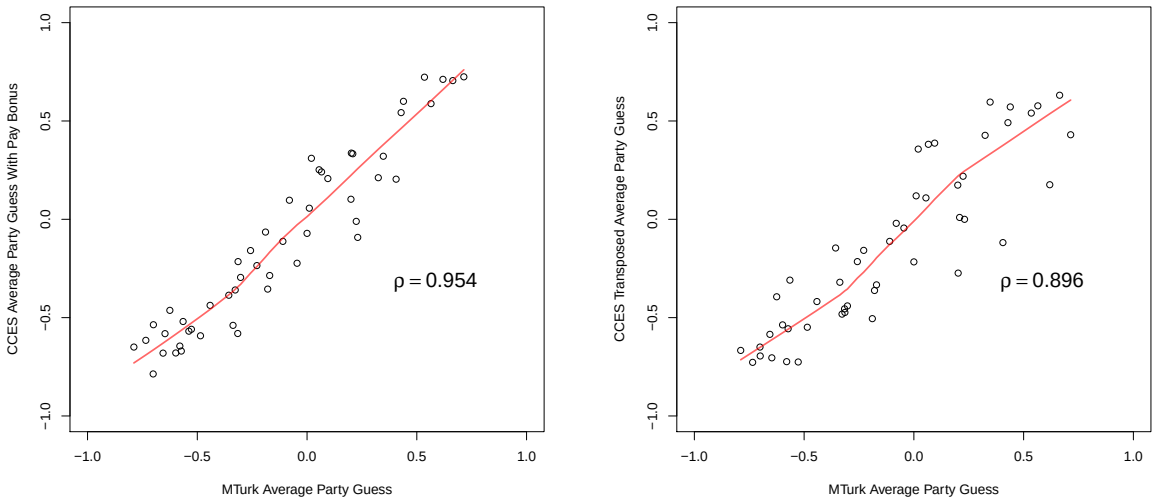
In the current MTurk study, I briefly explore the difference uncovered above between positive and negative frames. It is possible in a survey context that respondents minimize their attention when guessing, especially for positive ads, but generally pay closer attention to advertising during the campaign. Positive ads may be less interesting, and so respondents just guess with less consideration than with negative ads. To heighten respondent attention and interest, as described above, I provide (a small) material incentive for making correct guesses. The expectation is that both negative and positive

guessing will improve in accuracy under a reward, but that the biggest gains will be for positive ads, since the baseline there is so low.

One clear effect from providing a reward for correct answers is that respondents are much less likely to respond as 'Not sure', with virtually everyone making some guess. Yet, the rate of correct guesses does not change for either positive or negative ads, when paying respondents to think more carefully about their responses. This can be seen in Figure 13(a), which shows the scatterplot of normal party guesses from *Frame C* on the x-axis, and the same ads as guessed when respondents are provided rewards in *Frame F* on the y-axis. The correlation between the scores across the experiments is $\rho = 0.95$. Incentivizing correct guess fundamentally does not improve people's ability to discern party in positive or negative ads.

A future extension to this project will explore the ideological and partisan signals contained in negative ads more fully. One conjecture is that positive language describing an issue could lead voters to see that issue as closer to their own views, while casting it in a negative light would do the opposite, leading respondents to view the issue as farther away from their attitudes. A way to address this is to transpose positive ads into negative language, and negative ads into positive frames as done the experiments in *Frame E*. The results for this are shown in Figure 13(b). Though guessing is a bit noisier when ads are transposed, otherwise these distributions are statistically identical at a correlation of $\rho = 0.90$. In other words, the tone of the *language* in the ad itself seems to have very little impact on party guessing. Negative ads, when transposed into positive language, are (almost) as easily guessed correctly as they are when analyzed in their native negative form. The same is true for transposing positive ads. A future extension will alter the policy language in positive and negative ads to make each more or less specific when discussing issues, among others.

Figure 13: Scatterplots of Guessing Experiments Rewarding Correct Guesses and Transposing Positive and Negative Ads



(a) Reward Condition

(b) Transposing Ad Tone