An Automated Method to Estimate Survey Question Bias

Aaron Russell Kaufman*

Department of Government

Harvard University

aaronkaufman@fas.harvard.edu

September 6, 2018

Abstract

Many survey researchers are interested in gauging public support for government policy, but there is strong evidence that a question's wording affects responses to it. I develop the first automated and scalable method to predict the magnitude and direction of the partisan bias a question's wording may impose on survey responses, and show using a series of survey experiments that it outperforms public opinion scholars in predicting that bias. Using a novel data set of almost one million survey questions from 1997 to 2017, I then examine trends in partisan survey question biases over time. I find that while questions related to economic issues are relatively unbiased, questions related to Barack Obama become steadily more conservatively biased from 2008 to 2017. Questions related to abortion and immigration are generally conservative, while questions related to healthcare and education are consistently liberal. Substantively, my results suggest that measurements of American public opinion are systematically biased; I discuss the implications of this result for democratic representation. Methodologically, this paper opens up new opportunities for studying ideology from text, and for improving survey methodology and measurement in public opinion.

*I am deeply grateful to Steve Ansolabehere, Ryan Enos, Adam Glynn, Gary King, Luke Miratrix, Kevin Quinn, Maya Sen, Arthur Spirling, Alex Storer, Robert Ward, Chris Warshaw, Ista Zahn, and participants at the 2014 Summer Meeting of the Society for Political Methodology for their helpful comments and criticisms. I am grateful to Institute for Quantitative Social Science, the The Eric M. Mindich Research Fund for the Foundations of Human Behavior, AWS in Education, and the Roper Center for their generous support of this project.

1 Measuring Public Opinion

Some scholars consider ideologically consistent opinions regarding government policy and the state of national affairs a prerequisite to a healthy democracy (Schattschneider 1960; Key 1966). However, a long history of scholarship in American politics has cast doubts on whether most U.S. voters hold such opinions, finding instead that their opinions are often superficial and unbound by overarching ideologies like liberalism or conservatism (Converse 1964; Zaller 1992; Achen and Bartels 2016). Across panel studies, respondents often show preference instability over time, answering in favor of a policy during the first wave, against it in the second, and in favor of it again in the third (Campbell et al. 1960 (though in the aggregate these issues may have coherenceAnsolabehere et al. 2008). Such instability impedes public officials' abilities to gauge their constituents' preferences, leading to policy outcomes inconsistent with them (Butler 2011).

Exacerbating this problem is that in order to gauge public opinion in the absence of revealed preferences, we rely on imperfect survey instruments. Surveys are prone to numerous biases, including survey nonresponse bias (Armstrong and Overton, 1977; Gelman et al., 2016), item nonresponse bias (Tourangeau and Yan, 2007), accessibility bias (Iyengar, 1990), social desirability bias (Grimm, 2010; Zizzo, 2010), priming bias (Podsakoff et al., 2003), framing bias (Tversky and Kahneman, 1985; Sniderman and Theriault, 2004; Entman, 2007; Krosnick et al., 2014), dog whistle bias (Albertson, 2015), and doubtless more yet to be discovered (Chong and Druckman, 2007).

The aggregate effect of these biases is that a respondent's answer to a survey question is endogenous to the survey instrument itself: survey respondents cannot consider a question independent of the surrounding survey or of their personal context. An especially pernicious form of bias is partisan framing bias, in which a question's framing induces partisan associations, thereby affecting survey responses[1]. For example, survey takers asked about abortion will respond differently if the prompt mentions religion than if it mentions healthcare; if asked to consider a question about a hot button issue, respondents may respond with the intention of satisfying researchers. And when considering a question with political cues written into the text, respondents may assess their political context before answering.

Partisan framing bias is important because it is both ubiquitous. Survey questions must be communicated using words, either written or verbal, and all language has intrinsic political valence (Lasswell, 1965; Wodak, 1989; Lakoff, 2010; Edelman, 2013). It is therefore impossible to write a survey question that everyone understands the same way. Thoughtful and experienced survey researchers may go to great lengths to mitigate these biases, but in many cases, research design choices reduce to tradeoffs between one form of bias or another.

Survey question-induced partisan biases may be ubiquitous and difficult or costly to mitigate, but they may still be largely inconsequential if their magnitudes are small. However, canonical examples illustrate how severe these biases may be. In a 2004 study of framing bias, Sniderman shows that 85% of respondents answered in favor of

[1]In this paper, I focus largely on partisan framing bias. For convenience, I refer to it as either partisan bias or survey bias interchangeably.

allowing a hate group to hold a political rally when the sentence was prefaced with "Given the importance of free speech", while only 45% of respondents answered as such when prefaced with "Given the risk of violence" (Sniderman and Theriault 2004). This 40 percentage point swing is remarkable, but Krosnick, Malhotra, and Mittal (2014) find similarly large variation from questions used to identify Barack Obama's birthplace, showing that both political preferences and empirical facts are subject to survey question bias.

While these academic examples may be contrived, a conscientious researcher seeking to gauge comparable preferences for abortion policy can select from more than 2,800 unique questions on that topic indexed by the Roper Center's iPoll Database, or choose to write their own; iPoll indexes more than 2,500 questions related to gun control, and more than 4,500 questions related to immigration. The variation between any two of these questions may be enormous, and leaves researchers with substantial freedom to select questions likely to produce results in line with their hypotheses or political agendas.

And while the implications for research are substantial, the potential aggregate affects of biased surveys, especially partisan biased surveys, are enormous. Not only does the process of taking surveys affect survey respondents' preferences[2] (Zwane et al., 2011), meaning that taking biased surveys affects preferences in partisan ways, but survey *results* affect public opinion and behavior in the aggregate (Ansolabehere and Iyengar, 1994; Boudreau and McCubbins, 2010; Kam and Utych, 2011; Utych and Kam, 2013; Rothschild and Malhotra, 2014; Gelman et al., 2016). If the bulk of the literature is correct that poll results affect voters' willingness to turn out, there exist strong incentives to produce biased polls and thereby improve election outcomes. By the same token, discrediting strongly biased polls and promoting less biased ones may improve the state of democratic representation.

Defining & Measuring Survey Bias

This paper begins with the hypothesis that posing a question with Republican-sounding language will produce more responses aligned with the Republican position as compared to a question contextualized or written in a Democratic or moderate style. I formalize this into a mathematical definition of relative survey bias, and I demonstrate that it is possible to predict the direction and magnitude of the survey bias induced by a particular question wording using a Wordscores model (Laver et al. 2003). While public opinion experts and survey researchers spend substantial time pretesting their surveys to qualitatively measure this bias, I show that the predictions generated by this model outperform those produced by political scientists and public opinion researchers, casting doubt on the survey industry's role as purveyors of objective measurement.

The rest of this paper proceeds as follows. In Section 2 I formulate a definition of survey bias. In Section 3, I introduce my measurement strategy for proxying the bias of arbitrary survey questions. In Section 4, I outline my experimental design in validating the measures produced by my model, and in comparing them to those produced

[2]The American National Election Study has sometimes been called "the most expensive voter education campaign in history."

by experts; I display the results of this analysis in Section 4. I then proceed in Section 5 to extend this measure to a data set of more than one million survey questions from 1994 to 2018, revealing trends in survey bias across issue areas and survey firms. Section 7 concludes with a discussion of future lines of inquiry related to this work as well as best practices in implementing my model and interpreting its results. Supplemental results and the web implementation of my survey bias model appear in the Appendix.

2 Defining Survey Bias

I first define a model of survey response based on the classic survey research literature and explain its statistical properties as in Imai et al. 2008. I illustrate by example how this framework ignores a fundamental property of survey research: there is no notion of unbiasedness in politics, and subsequently no unbiased survey question. Incorporating this property, I refine a new survey research framework and derive unbiased estimators for its quantities of interest, such as population average response, and a new quantity of interest, *relative* survey question bias.

Classic Survey Research

An individual i's response to a survey question q, given by $y_{i,q}$ is a function of that individual's true preference θ_i plus that survey question's bias β_q, which is stable for all respondents, and an error term $\epsilon_{i,q}$:

$$y_{i,q} = \theta_i + \beta_q + \epsilon_{i,q} \tag{1}$$

In a classic survey study, the quantity of interest is the population average preference $\theta = \frac{1}{N}\sum_{i=1}^{N}\theta_i$. However, we observe only a sample of n responses to a survey question $y_{i,q}$ from the total population N; typically, $n \ll N$. Define three *sample average quantities*:

1. $\bar{y}_{i,q} \equiv \frac{1}{n}\sum_{i=1}^{n} y_{i,q}$
2. $\bar{\theta} \equiv \frac{1}{n}\sum_{i=1}^{n}\theta_i$
3. $\bar{\epsilon_q} \equiv \frac{1}{n}\sum_{i=1}^{n}\epsilon_{i,q}$

Then show:

Lemma 1

$$
\begin{aligned}
\bar{y}_q &= \frac{1}{n}\sum_{i=1}^{n} y_{i,q} \\
&= \frac{1}{n}\sum_{i=1}^{n} (\theta_i + \beta_q + \epsilon_{i,q}) \\
&= \frac{1}{n}\sum_{i=1}^{n} (\theta_i) + \frac{1}{n}\sum_{i=1}^{n} (\beta_q) + \frac{1}{n}\sum_{i=1}^{n} (\epsilon_{i,q}) \\
&= \bar{\theta} + \beta_q + \bar{\epsilon_q}
\end{aligned}
$$

In order for $\bar{y}_q$ to be an unbiased estimator of the quantity of interest θ, we make three assumptions:

A1: Random sampling[3]
A2: $\beta_q = 0$
A3: $E[\bar{\epsilon}_q] = 0$

Given these three assumptions,

$$
\begin{aligned}
E[\bar{y}_q] &= E[\bar{\theta} + \beta_q + \bar{\epsilon_q}] \\
&= E[\bar{\theta}] + \beta_q + E[\bar{\epsilon_q}] \\
&= \theta
\end{aligned}
$$

Quantities of Interest in a Relative Bias Framework

This classic formulation, while foundationally important, makes some unverifiable, and some impossible, assumptions. I recharacterize this model, and its parameters. Doing away with the notion of survey bias and fundamental error as having meaningful zero points allows me the flexibility to estimate alternative quantities of interest.

In the classic survey literature, $\bar{\theta}$ indicates a sample average underlying preference for a policy, which is equal to the average survey response in the absence of bias. However, in politics, it is axiomatic that there is no neutral, unbiased position, and by extension, there is no neutral or unbiased survey question; therefore it is nonsensical to search for one. Instead, I consider the relative effects of one biased survey question versus another, and I thus recharacterize $\bar{\theta}$ as a reference point, and β_q as a relative deviation from that reference point. Therefore, to assess a survey question's bias, I must do so relative to another question. I select as my quantity of interest $\tau_{q_1-q_2}$, the relative bias of q_1 compared to q_2, which I define as:

$$
\hat{\tau}_{q_1-q_2} \equiv \beta_{q_1} - \beta_{q_2} \tag{2}
$$

To understand why I select this quantity of interest, consider a researcher deciding which of two survey questions, q_1 or q_2, to field. Both questions seek to gauge support for the same policy, but do so with different language. The

[3]For example, the sample is 1 to n of $1, ..., n, ..., N$ where the indices of N are randomly permuted.

researcher may decide to pilot both versions of the study: half of the pilot respondents will receive q_1, and the other half, q_2. Under randomization, we expect both sample populations to have the same preferences $\bar{\theta}$. Therefore, it must be the case that the difference in observed outcomes, $\bar{y}_{q_1} - \bar{y}_{q_2}$, is the difference in survey bias induced by q_1 *relative to* q_2. Our researcher, noting minimal differences, may choose to proceed with either question; if instead, she finds a large difference, she may undergo additional pretesting, or select a question likely to produce more conservative estimates of her expected effects later in the study.

There are two main takeaways from this illustration. The first is that to identify the quantity of interest $\hat{\tau}_{q_1 - q_2}$, I must hold constant $\bar{\theta}$ while varying the survey question used to solicit $\bar{y}_q$. One plausible experimental design is to assign any one respondent both q_1 and q_2 in randomized order, though this design would induce spillover effects and artificially decrease the observed difference between β_{q_1} andβ_{q_2}. Instead, I randomize assignment of q_1 and q_2 among a survey sample. Any subsequent difference between $\bar{y}_{q_1}$ and $\bar{y}_{q_2}$ can only be caused by the survey question and not by differences in underlying preferences θ between the respondents assigned to q_1 and those assigned to q_2.

Estimators

The second takeaway is a straightforward estimator for $\hat{\tau}_{q1-q2}$:

$$\hat{\tau}_{q1-q2} \equiv \bar{y}_{q_1} - \bar{y}_{q_2} \tag{3}$$

To prove the unbiasedness of this estimator, I first note the following equalities from Lemma 1:

$$\bar{y}_{q_1} = \bar{\theta} + \beta_{q_1} + \bar{\epsilon}_{q_1}$$
$$\bar{y}_{q_2} = \bar{\theta} + \beta_{q_2} + \bar{\epsilon}_{q_2}$$

Subtracting the bottom equation from the top yields:

$$\begin{aligned} \bar{y}_{q_1} - \bar{y}_{q_2} &= \bar{\theta} + \beta_{q_1} + \bar{\epsilon}_{q_1} - \bar{\theta} - \beta_{q_2} - \epsilon_{q_2} \\ &= \beta_{q_1} + \epsilon_{q_1} - \beta_{q_2} - \bar{\epsilon}_{q_2} \end{aligned}$$

We find that our estimator in expectation yields our quantity of interest from Equation 2:

Lemma 2

$$
\begin{aligned}
E[\bar{y}_{q_1} - \bar{y}_{q_2}] &= E[\beta_{q_1} + \bar{\epsilon}_{q_1} - \beta_{q_2} - \bar{\epsilon}_{q_2}] \\
&= E[\beta_{q_1}] - E[\beta_{q_2}] + E[\bar{\epsilon}_{q_1}] - E[\bar{\epsilon}_{q_2}] \\
&= E[\beta_{q_1}] - E[\beta_{q_2}] && \text{Assumption A3} \\
&= \beta_{q_1} - \beta_{q_2} \\
&= \hat{\tau}_{q_1 - q_2}
\end{aligned}
$$

Modeling

This estimator illustrates a valuable way to capture the relative bias induced by a pair of survey questions. By randomly assigning one question versus another, researchers can obtain unbiased estimates of the relative bias of those questions. However, this design becomes intractable very quickly as one increases the number of survey questions, and statistically distinguishing between as few as five survey questions can require several thousand survey respondents.

In the following section, I propose to use a nonparametric text-based model to estimate the relative bias of survey questions without relying on survey respondents, and then in Section 4, I show that predictions generated by this this model largely agree with the unbiased estimates produced by a gold-standard survey experiment described in Section 2. This text-based model scales efficiently and holds great promise for aiding survey researchers; I provide a web-based implementation of this model in Appendix A.

3 Data, Measurement, & Methodology

The methodological challenge in this paper is to estimate the relative bias induced by survey questions used to gauge political preferences. To capture this partisan dimension of a survey question, I first train a simple supervised learner, Wordscores, to model the relationship between partisanship and word usage in Congressional text. I then use this model to predict the partisan bias of survey questions[4]. I collect three sources of data: (1) Congressional press releases and floor statements, (2) Congressional ideal points, and (3) a corpus of survey questions. The Congressional press releases and floor statements serve as training documents. Each press release or floor statement I associate with its author's DW-NOMINATE first dimension; these ideal points serve as training labels to construct a statistical model of partisan bias as a function of word usage. The corpus of survey questions, drawn from the Roper Center's iPoll database, compose the set of documents whose relative bias I aim to estimate.

I follow a standard methodological procedure in the Text as Data literature to produce a trained text partisanship model and to estimate the relative bias of the survey questions in my corpus. The method consists of five steps: (1) Collect an unstructured training corpus which varies along the dimension of interest, (2) label each document in the

[4]Recent work at the intersection of causal inference and text analysis (see, for example, (Roberts et al., 2013; Fong and Grimmer, 2016) and (Egami et al., 2018)) characterizes this problem as a text-as-treatment problem: the dependent variable, in this case average survey response, is a function of a text, the survey question.

corpus corresponding to its location along the dimension of interest, (3) convert the unstructured training corpus to a structured data set, (4) train a supervised model on the training corpus and training labels, and (5) use the model to predict the labels for out-of-sample documents, the ultimate quantities of interest. I detail these five steps below.

1. Building training and test corpora First, I construct a data set of everything spoken on the floor of the Senate, the House of Representatives, or in Conference Committees from January 1994 to June 2017, and every press release produced by a Member of Congress in this same interval. This corpus includes nonpartisan text such as procedural speech and speeches in memorial to important constituents, as well as highly ideological rhetoric including One Minute Speeches and floor debates over proposed legislation. From these Congressional texts I create a data set where each document consists of continuous speech by exactly one individual along with associated metadata including date and speaker. This corpus includes more than 8 million documents with over 20,000 unique unigrams and bigrams common to at least 1% of documents, and encompasses text from every member of Congress during this period. Next, I compile a corpus of every survey question in the Roper Center's iPoll Database[5] asked from January 1994 to June 2017. Associated metadata includes the dates in which the survey was fielded, the organization which fielded the survey, and topic area indicators. This set of survey questions comprises my test set, the documents for which I aim to estimate the partisan bias.

2. Attaching training labels In the second step, I apply training labels to each document. I match each press release or Congressional statement to its author's DW-NOMINATE first dimension (Carroll et al. 2009) *during the year of its authorship* as a measure of that document's partisanship. Documents co-authored by multiple Members of Congress (for example, joint press releases) receive as their training labels the average of their authors' DW-NOMINATE first dimensions.

3. Processing a corpus into a data set To convert these documents into analyzable features, I rely on the "bag of n-grams" model[6], which views a document as an unordered collection of the words which comprise it. I convert all documents into a series of unigrams, single words, and bigrams, an ordered pair of words, and their associated frequencies within documents. The phrase "universal healthcare" becomes three variables: the unigram "universal", the unigram "healthcare", and the bigram "universal healthcare", each with a frequency of 1. This corpus representation is called a "term-document matrix". I produce two such matrices: a training matrix, derived from the press release and Congressional floor speech corpora, and a test matrix, derived from the iPoll survey question corpus. There are numerous preprocessing decisions in this step: I use term frequency-inverse document frequency weights, preserve stopwords, and remove terms that appear in fewer than 1% of all documents (Denny and

[5] http://www.ropercenter.uconn.edu/CFIDE/cf/action/ipoll/

[6] There exist within the Natural Language Processing literature an enormous collection of models capable of extracting more information from a document, including Brown clustering (Brown et al. 1992) and paragraph vectors (Le and Mikolov 2014). I rely on the bag of n-grams model for simplicity, interpretability, computational tractability.

Spirling, 2018).

4. Training a Wordscores model In the fourth step, I train a Wordscores model (Laver et al., 2003) in which the dependent variable is the vector of training labels from Step 2 and the covariate matrix is the training term-document matrix from Step 3. For each term in the term-document matrix, the Wordscores model produces a point estimate and confidence interval indicating where on the DW-NOMINATE scale this word lies. A substantive interpretation for this estimate is that for a given word w and word score $s(w)$, the expectation of the author's DW-NOMINATE score $s(w)$. Wordscores offers many desirable properties for this application. First, it induces no sparsity. Most machine learning models used in text applications regularize coefficients to produce parsimonious models. However, in estimation problems using short documents like survey questions, it is valuable to estimate bias for every lexical feature rather than the most discriminating. Second, Wordscores is computationally efficient (Benoit and Nulty, 2016), allowing interested researchers to compute predictions for their own questions quickly. Third, it encodes no functional or distributional assumptions (Lowe, 2008). For a detailed discussion of Wordscores, see Appendix C.

5. Generating out-of-sample predictions In the fifth step, predicting the document scores for out-of-sample survey questions, I use the trained model from Step 4 to estimate the survey bias of every question included in the test corpus from Step 3. Mathematically, the prediction step averages the calculated word scores corresponding to the text of each survey question, weighted by each word's frequency. The survey question "Do you support or oppose providing a legal way for undocumented immigrants already in the United States to become US citizens?" has a document score of -0.214, the weighted average of a series of individual feature scores in Table 1 (for details, see Appendix C). We find that "undocumented" and "immigrant" both strongly indicate a liberal question, while "way" and "become" both indicate conservative questions.

Table 1: Word Scores for the survey question: "Do you support or oppose providing a legal way for undocumented immigrants already in the United States to become US citizens?" Words more used by Democrats are negative, while words more used by Republicans are positive. The estimated partisanship of this question -0.214, a strongly liberal score.

	Word	Word Score
1	away	0.0566
2	already	0.0089
3	become	0.1182
4	immigrant	-0.1165
5	oppose	0.0227
6	provide	0.0333
7	state	0.0081
8	support	-0.0408
9	undocumented	-0.2103
10	united	0.0119
11	united states	0.0014
12	way	0.1011

4 Modeling & Substantive Validity

I perform two studies to validate the performance of my model. The first tests internal validity: Can the model reliably and accurately estimate the partisanship of a document produced by a Member of Congress? The second study tests external validity: Do survey respondents in the aggregate respond to survey questions in a manner consistent with my model's predictions?

For the first test, I conduct a standard cross-validation analysis and show that my model can reliably identify documents as either written by Republicans or written by Democrats. In the second study, I ask survey respondents to express their support for various proposed policies using randomly assigned survey questions, showing that my model identifies survey question bias more accurately than public opinion experts asked to rank order those same questions.

Internal Validity

As a measure of this model's internal validity, I perform 10-fold cross-validation (Arlot et al., 2010) to calculate its predictive accuracy. Predictive accuracy is a necessary condition to this exercise: if my model cannot capture the distinguishing characteristics of liberal versus conservative speech among Members of Congress, it certainly cannot do so among survey questions. By showing that my model can with high precision locate Members of Congress on the ideological spectrum using the text of their writings, I can claim with confidence that the model captures characteristics of partisan text.

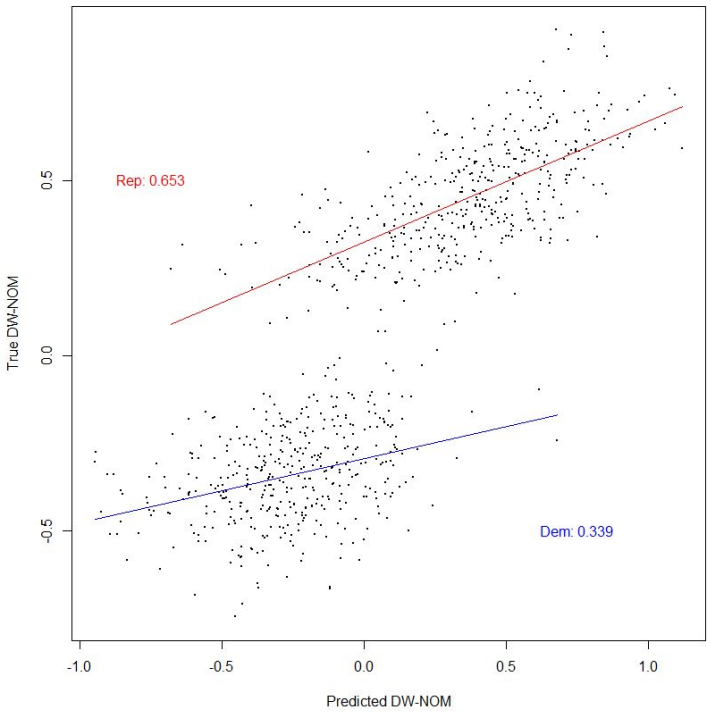
Generally, a cross-validation procedure involves partitioning a training set into a number of non-overlapping random subsamples, training a model on all but one of them, predicting the outcome measure for the held out subsample, and comparing the model's predictions to the true outcome labels. If the predictions or correlate highly with the training labels, then the model is assessed as valid and its predictions on the training set may be used with confidence. In this case, the unit of analysis is the Member of Congress: I collapse each Member of Congress' text for one year into a single document, then partition the data set into ten random samples[7]. Each fold consists of approximately 90 member-years.

A cross-validation scatterplot is in Figure 1. Each point indicates one year for one Member of Congress from 1994 to 2017; the Y-axis indicates each Member's DW-NOMINATE first dimension, while the X-axis indicates that Member's ideology estimated from their text corpus. My model successfully identifies the party ID of a document's author in 86% of cases. Within Republicans, the model's predicted ideology for a document i correlates with that document's author's DW-NOMINATE first dimension at 0.653, while among Democrats, it correlates at 0.339. As well, the model correctly classifies Democrats as being either liberal, mainstream, or centrist Democrats with 62.9%

[7]During the random partition process, I ensure that no Member of Congress appears in more than one partition; otherwise, the model might learn to identify the particular speech patterns of individual *members* rather than the characteristics of their ideology.

accuracy, and correctly classifies Republicans as being conservative, mainstream, or centrist Republicans with 67.2% accuracy.

Figure 1: Using ten-fold cross-validation, I estimate that my Wordscores model correctly identifies a Member of Congress's party based on their press release text with 86.2% accuracy.



The model's results coincide with recent evidence that there is more ideological variation among Republicans than among Democrats, and importantly, that variation within a party manifests more clearly in text and speeches among Republicans than among Democrats (Russell, 2017).

External Validity

There are innumerable potential survey questions to gauge preferences for any given policy, ranging from liberally biased to conservatively biased. If we could gather an unlimited pool of survey respondents and assign to each one a random question along that spectrum, we could order each question according to its aggregate response, and therefore its bias. Pollsters and public opinion researchers may run a subset of this experiment (pretesting), or may perform this ranking speculatively, but for them to systematically rank a large set of questions would be prohibitively time-intensive. A successful text-based model of survey question bias should be able to correctly order each of these questions according to their bias without relying on an unfeasible experiment, and without unnecessarily burdening, as well or better than survey researchers and public opinion scholars.

I conduct and analyze two surveys to measure whether survey research experts or my Wordscores model can more accurately rank survey question bias. In the first survey, I randomly assign a series of questions to respondents on Mechanical Turk (Berinsky et al., 2012; Goodman et al., 2013) and Harvard DLABSS (Enos et al., 2016) as in the gold standard experiment in Section2, and the second asks survey research experts to rank order a series of survey questions according to their bias[8]. Viewing the rank ordering produced by the gold standard experiment as the ground truth, I then compare the ranking produced by my model to the ranking produced by aggregated survey researcher experts, and find that my model's ranks agree more closely.

I wrote 11 questions across four issue areas – three about gun control, and four each about marijuana legalization and about accepting refugees. Within an issue area, each question shares a common syntactic structure and word choice, though each question includes a short framing clause. All questions take the form of a statement, to which respondents are prompted to indicate their level of agreement on a seven-point scale: Strongly Agree, Agree, Somewhat Agree, Neither Agree nor Disagree, Somewhat Disagree, Disagree, Strongly Disagree. Generally, questions are of the form: "Considering/Despite [consideration], do you support [policy]?"

For the first study, I recruited 1000 respondents, 750 from Mechanical Turk and 250 from the Harvard Digital Laboratory for the Social Sciences, or DLABSS[9], in March and April 2017. Respondents received one question each from the three issue areas. Within each, respondents received a random question drawn from the pool of three to four questions[10]. I code survey responses from 1 to 7, where 1 is the most liberal response and 7 is the most conservative, and calculate the *average* response for each question. Due to random assignment, as I show in Section 2, I can interpret the difference in average response between questions in the same issue area as the treatment effect of asking one survey question rather than another.[11] Importantly, I note that the rank orders produced by the Mechanical Turk sample match precisely the rank ordered produced by the DLABSS respondents, providing valuable evidence that the obtained relative bias scores are reliable.

For the second study, I recruited 12 political scientists and public opinion scholars and asked them to rank the survey questions, within issue area, from most liberal to most conservative. To analyze this study, I calculate the average rank for each question where 1 is the most liberal and 3 (or 4) is the most conservative. Contrary to the previous study, the expert survey respondents had low reliability, with any one expert often correlating negatively with other experts.

[8] In this second survey, the experts were told that I intended to conduct a survey experiment to measure survey question bias among Mechanical Turk respondents.

[9] Approved by Harvard IRB, submissions IRB14-2617 and IRB16-0530

[10] I randomized issue area order to mitigate spillover framing.

[11] Rather than reweighting respondents to emulate a nationally representative sample and estimate a population average treatment effect, I estimate a sample average treatment effect and ask my expert panel to estimate effects among that same population. In practice, the PATE and SATE are unlikely to be very different (Miratrix et al., 2017).

Question Text	Revealed Rank	Model Rank	Expert Rank
We should require mental health checks before anyone can purchase a firearm.	1	**1**	**1**
Even considering the importance of protecting oneself, we should require background checks before anyone can purchase a firearm.	2	**2**	**2**
Even considering the importance of the Second Amendment protecting the right to bear arms, we should require background checks before anyone can purchase a firearm.	3	**3**	**3**
Considering the importance of states' rights, marijuana should be legal to prescribe for approved medical patients.	1	**1**	4
Considering the economic benefits, marijuana should be legal to prescribe for approved medical patients.	4	**2**	**2**
Despite the risks to public safety, marijuana should be legal to prescribe for approved medical patients.	3	**3**	**3**
Despite federal law, marijuana should be legal to prescribe for approved medical patients.	2	4	**1**
Considering the economic benefits of immigration and cheaper labor, we should accept more refugees into the US.	1	**1**	4
Considering the importance of helping to avert a humanitarian crisis, we should accept more refugees into the US.	4	**2**	1
Despite the threat to national security, we should accept more refugees into the US.	3	**3**	**3**
Despite the high unemployment rate, we should accept more refugees into the US.	2	4	**2**

Table 2: The Wordscores model predicts survey question bias as well or better than aggregated expert ranks. For each question, the closer rank is bolded; both ranks are bolded in the case of ties.

Experimental Results

In Table 2, I present the results from these two surveys and my modeling analysis: the Mechanical Turk average response rank, the average expert rank, and the Wordscores predicted rank. Experts do no better than the Wordscores rank: my model outperforms the experts in the immigration and medical marijuana cases; both models tie in the gun control case.

The correlation in ranks between my model and the revealed rank is 0.691; between the experts and the revealed rank, it is -0.159, for a difference in correlations of 0.850. To assess how likely it is that this difference arises by chance, I conduct a permutation test in which I randomly shuffle both my model's and the experts' rankings, then calculate the difference in correlations. The observed difference is larger than 99% of random draws, yielding a p-value of approximately 0.01.

While *aggregations* of expert opinions fail to outperform the Wordscores model, expert opinions have a second drawback: inconsistency. Each expert produces a ranking of questions within each issue area, and I correlate each set of ranks, then take an average of all correlations to measure intercoder reliability as in Kaufman et al. (2017).

The gun control ranks correlate at 0.003; marijuana ranks correlate at 0.055; and refugee question ranks correlate at -0.033. To address concerns that the measurement strategy I use to gauge expert opinions precludes strong intercoder reliability, a placebo test using abortion questions, illustrated in Appendix B, produces intercoder correlations of 0.358; my sample of experts simply disagreed to a high degree.

I have shown that the Wordscores model performs at least as well as expert researchers in the aggregate, without the drawback of human inconsistency, and at scale with computational efficiency. However, it is important to be cautious when examining the bias of any single pair of questions, especially similar ones, as measurement error and modeling error produce relatively large uncertainty estimates around any single relative bias score. As a result, in the following section I examine trends in survey question bias across hundreds of thousands of survey questions, thereby mitigating the measurement error problem.

5 Survey Firm & Issue Area Bias

If it is the case that among some issue areas survey questions are consistently biased in either a liberal or conservative direction, we may be collecting biased estimates of public opinion, which may hold implications for representation and legislative outcomes (Butler, 2011). To this end, I collected every survey question indexed by the iPoll Database at the Roper Center from January 1994 to August 2018, along with the dates on which the questions were fielded, the firm which wrote the question, and the issue areas under which each question falls[12]. This totals 430,929 unique questions across 2,109 unique surveyor organizations. Then, using my Wordscores model, I estimate the bias for each of these questions. Note that in training the Wordscores model, I use the entire training set at once. That is, survey questions written in 1994 or 2018 are scored in reference to partisan word usage averaged over the time period of 1994 to 2018.

In the analyses below, I compare aggregated survey questions across firms and over time, relative to certain relevant baseline values, so as to produce valid comparisons with a meaningful zero-point. In examining the trend in aggregate bias within an issue area or a firm, I use as a baseline that issue or firm's January 1994 average score, subtracting that value from all subsequent bias estimates. In comparing *across* issue areas or firms, I use as a baseline the average score of all survey questions in the corpus.

House Effects in Survey Bias

In Figure 2, I present the average bias from 1994 to 2017 for six of the largest survey-fielding organizations: CNN, NBC, ABC, CBS, the New York Times, and the Wall Street Journal; these bias scores are all calculated relative to

[12] Surveys may be produced by a single firm, like a "US News & World Report Poll", or may be produced by multiple organizations, for example a "Gallup/CNN Poll" or a "Kaiser/Los Angeles Times Poll". Gallup, for example, has published polls with more than 10 other firms since 1994. To accommodate this, I treat a survey question from a "Gallup/CNN Poll" as one question from a Gallup and another from CNN (rather than a fractional observation).

a baseline of the average of all survey questions fielded in January 1994, the beginning of the time series. Together, they account for 145,930 survey questions over 22 years. All six firms cluster tightly around the baseline and vary only little, though all six begin to trend in a liberal direction in 2012. Then, in Figure 3, I show the average bias for each of 26 survey-writing firms relative to the average bias of all survey questions. Fox News and Reuters produce the most conservative questions, while Yale University, Harvard University, and Kaiser produce the most liberal questions. Kaiser, a healthcare provider, primarily fields questions related to healthcare. The Associated Press, Harris Poll, and Pew produce the most relatively unbiased questions.

Figure 2: Six major survey-producing organizations are all largely unbiased relative to a January 1994 baseline. Each point is a single survey question fielded by the firm indicated above each plot. The dashed grey line at $y = 0$ is for benchmarking only. The blue line represents a loess fit of each firm's relative partisan bias from 1994 to 2017.

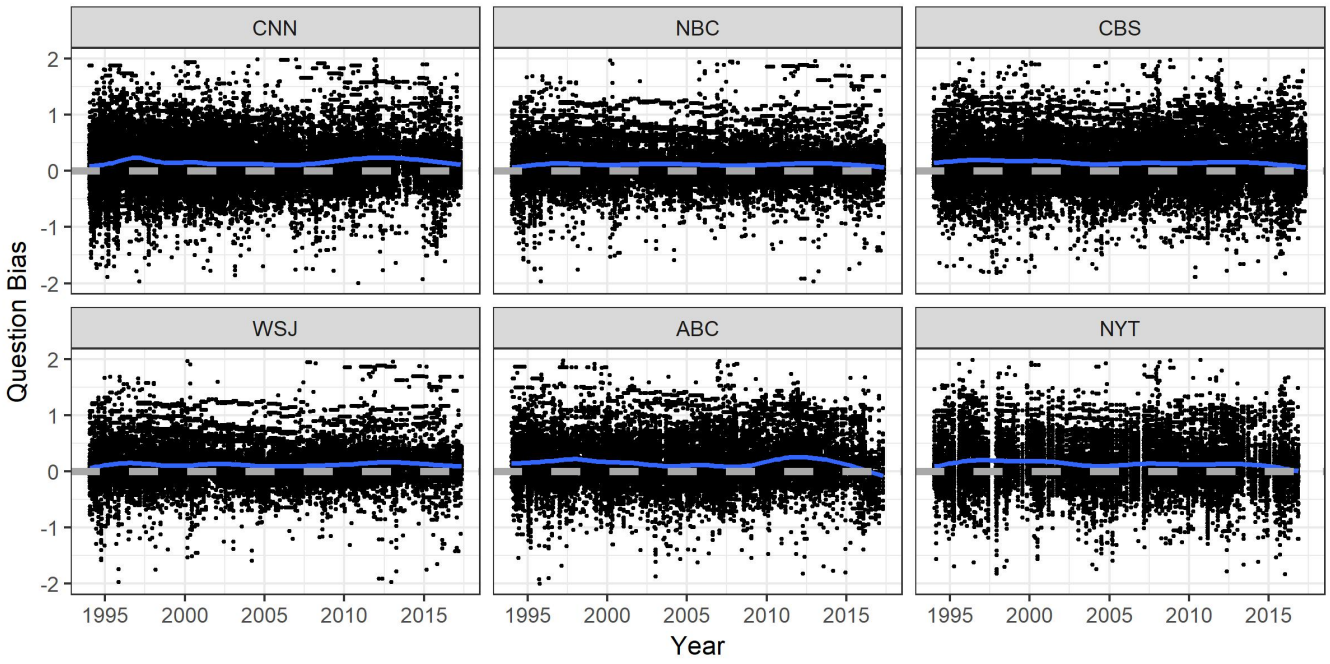
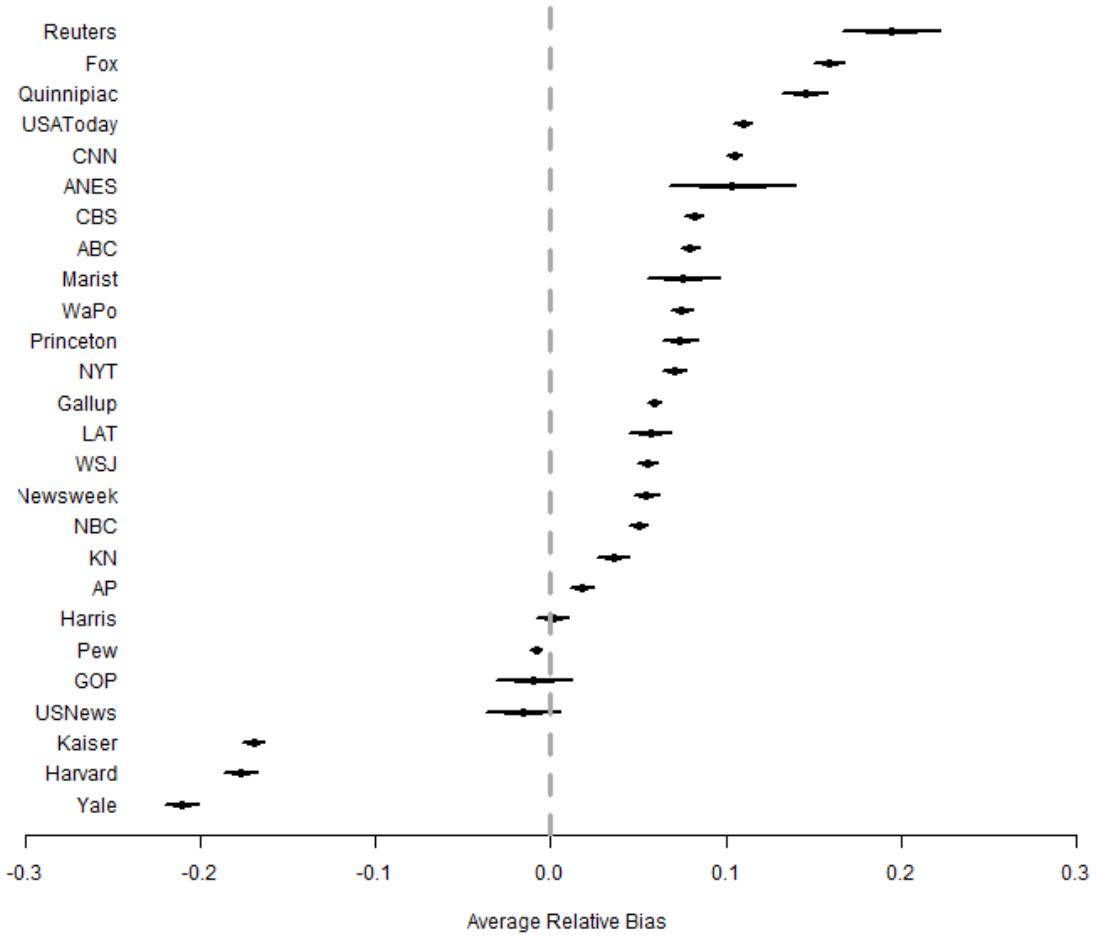


Figure 3: Most survey-producing firms produce conservatively biased questions relative to the average of all survey questions. Each point represents the average bias, with 66% and 95% confidence intervals, for all survey questions fielded by the corresponding organization. Fox News and Reuters produce the most relatively conservative questions, while Yale University, Harvard University, and Kaiser produce the most relatively liberal questions from 1994 to 2017.



Were Surveys Biased In Favor of Barack Obama?

In September and October 2012, Daily Kos and SEIU produced a survey showing that a plurality of all voters, and nearly three-quarters of Republicans, believe that surveys were biased in favor of then-president Barack Obama (Sink, 2012). To address these claims, I examine time trends in the aggregate bias of questions related to Barack Obama. Using pre-inauguration survey questions from 2007 as a baseline, Figure 4 shows exactly the opposite: among questions referencing Barack Obama, all 6 firms trend from ask increasingly *conservative* questions from 2008 to 2016. In 2016, NBC and the Wall Street Journal trend back to 2007 levels, while the remaining firms do not.

As a comparison, Figure 5 shows that 6 major survey firms, questions about the economy have remained static

over time, with only slight variation, despite smaller sample sizes. Similarly, questions about George W. Bush remained relatively stable over his tenure as well, shown in Figure 6.

Figure 4: Six major survey-producing organizations all produce systematically conservative questions regarding Barack Obama relative to their 2007 average; while NBC and the Wall Street Journal moderate those questions starting in 2016, the other four firms do not. Each point is a single survey question fielded by the firm indicated above each plot. The dashed grey line at $y = 0$ indicates the average of all questions related to Obama in 2007, while the blue line represents a loess fit of each firm's average partisan bias from 2007 to 2017.

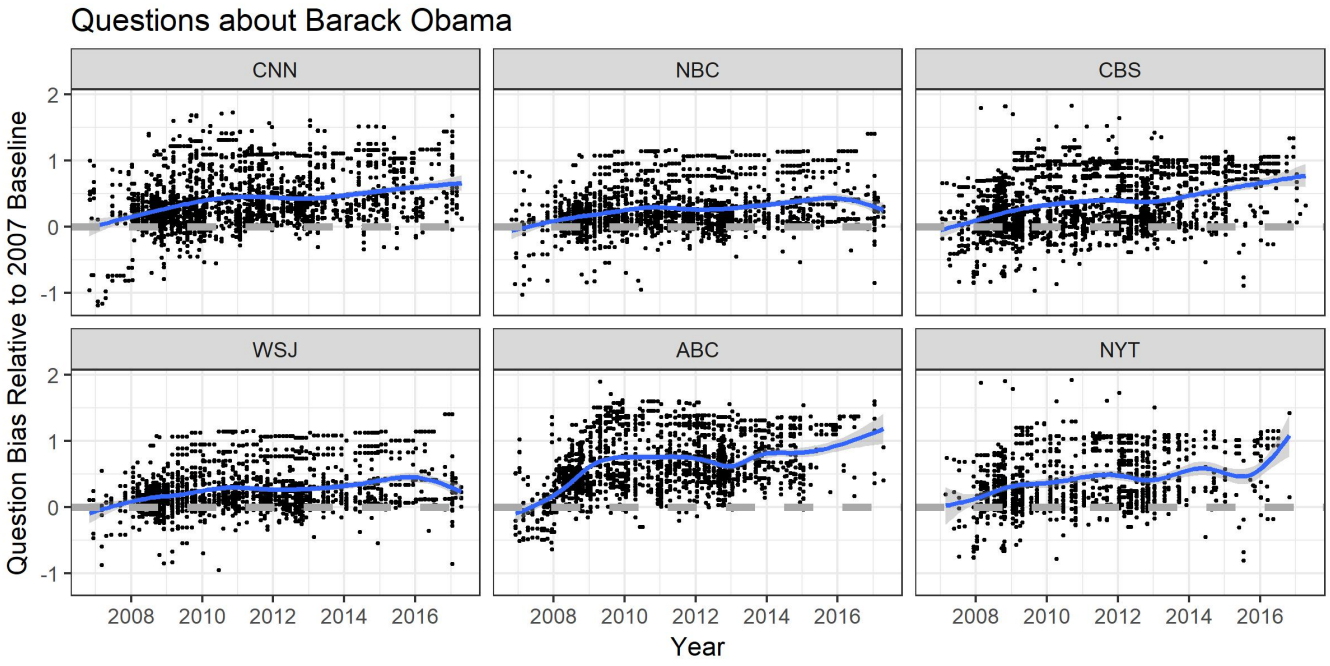


Figure 5: Six major survey-producing organizations are all consistent over time in discussions of the economy. Each point is a single survey question fielded by the firm indicated above each plot. The dashed grey line at $y = 0$ indicates the average bias of all questions related to the economy in January 1994, while the blue line represents a loess fit of each firm's average partisan bias from 1994 to 2017.

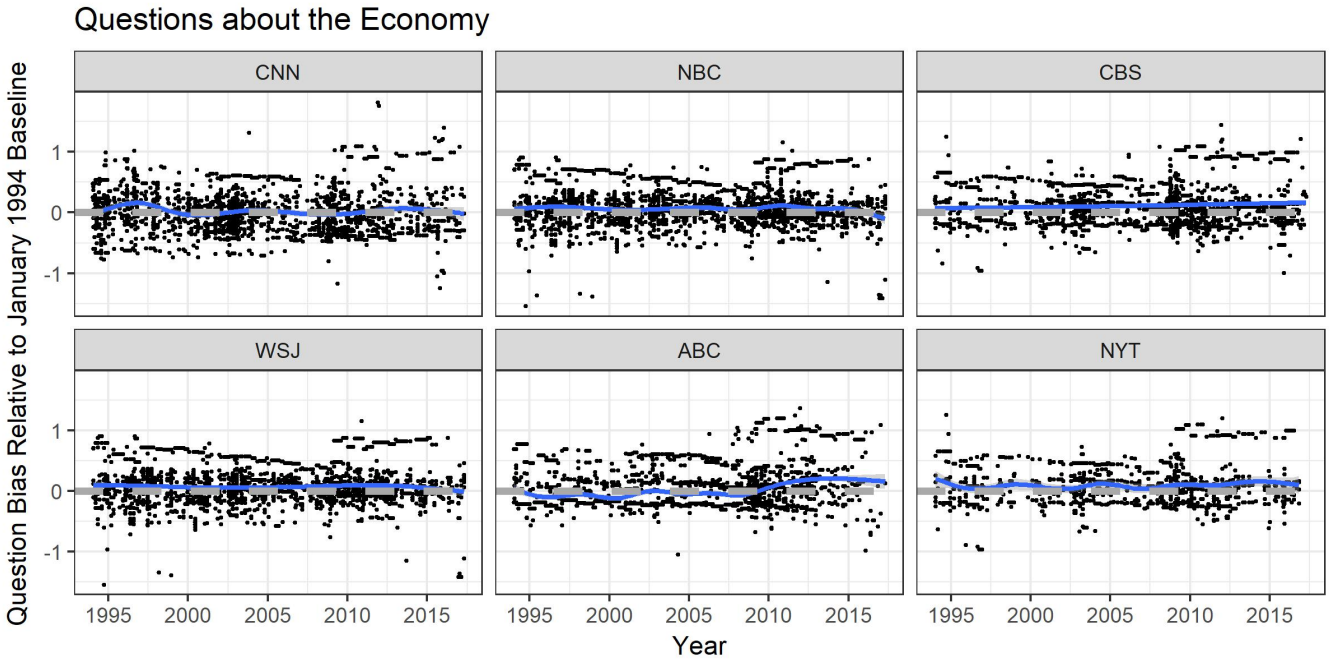
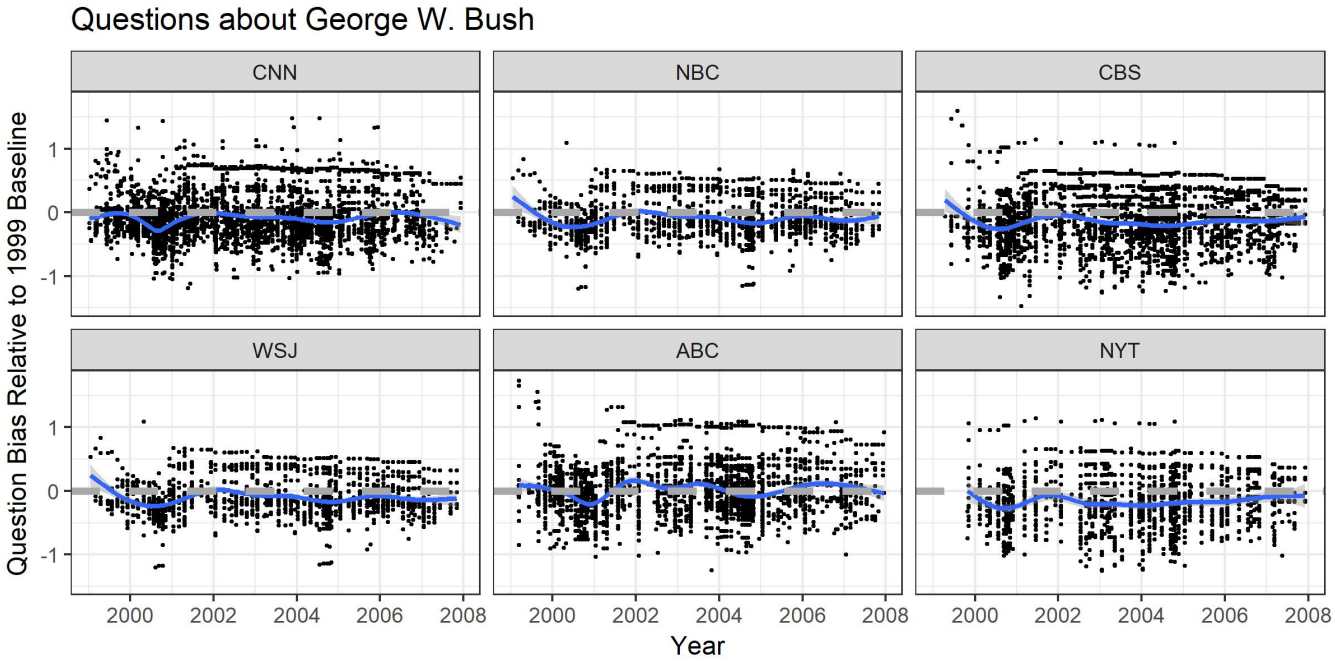


Figure 6: Six major survey-producing organizations are all consistent over time in questions related to George W. Bush. Each point is a single survey question fielded by the firm indicated above each plot. The dashed grey line at $y = 0$ is the average bias for questions related to Geroge W. Bush in 1999, while the blue line represents a loess fit of each firm's average partisan bias from 1999 to 2008.



This trend is worrying: if questions used to assess support for Barack Obama became increasingly conservative from 2008 to 2016, then perhaps perceptions of public support for Obama were skewed.[13] More concerning is that surveys drive public opinion as much as they measure it. A vast literature from psychology, marketing, public opinion, and political science indicates that the public tends to shift its opinions in line with survey results (Lang and Lang, 1984; Skalaban, 1988; Morwitz and Pluzinski, 1996; Areni et al., 2000; Sonck and Loosveldt, 2010; Rothschild and Malhotra, 2014; Roy et al., 2015), giving survey organizations substantial power to sway public discourse by fielding biased polls.

Issue Ownership & Survey Bias

To interpret trends in issue area bias over time, I turn to the issue ownership literature. Petrocik 1996; 2003; 2008 outlines a theory of partisan issue ownership, in which candidates and parties selectively emphasize the issues on which they are the strongest, and in doing so, align themselves to voters who consider those issues the most important.

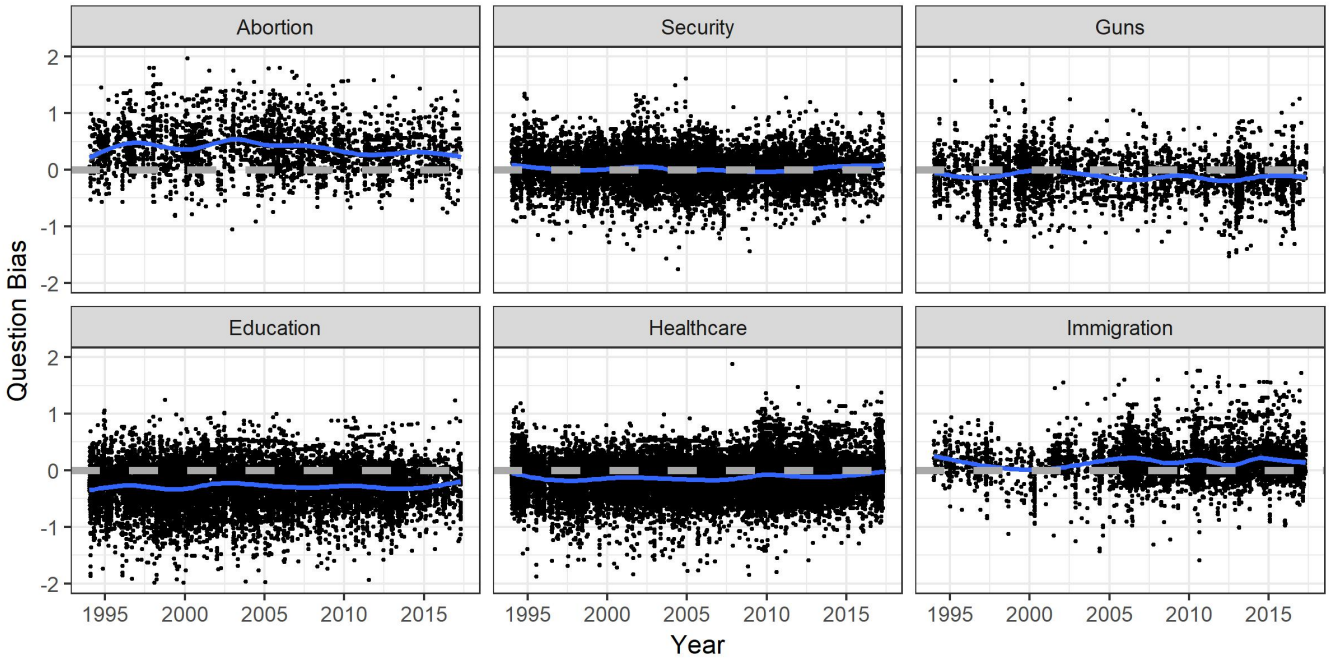
A straightforward implication of this theory is that an issue that is owned by the Republicans will see more discussion by Republicans in both Congressional floor speeches and in press releases; consequently, words related to

[13] I present a series of robustness checks to these results in Appendix D.

issues owned by Republicans will be scored as Republican by my model. By examining the aggregate bias in survey questions related to a given issue area, we can identify which party owns it, if any. However, using aggregate survey bias to study issue ownership provides additional inferential leverage: these predictions can shed light on both the temporal stability of issue ownership (Seeberg, 2017) and the relative magnitudes of owned issues.

In Figure 7 is a plot of six important political issues and their average bias over time, relative to the mean of all survey questions. Some topics, such as healthcare and education, are more common than others, while questions related to immigration become much more common in 2008. Looking at average bias, abortion questions are consistently conservative, while education questions are consistently liberal. This evidence is consistent with the issue ownership literature in that education and healthcare are owned by Democrats; however, contrary to that literature, security and gun control seem to be equally prevalent among both parties. Interestingly, immigration became increasingly the purview of Republicans beginning in the mid-2000s.

Figure 7: Each plot includes the universe of fielded survey questions, across all firms, related to a single topic. Each point is a single survey question related to the topic indicated above each plot. The dashed grey line at $y = 0$ indicates the average of all survey questions, while the blue line represents a loess fit of each topic's average partisan bias from 1994 to 2017.



6 Best Practices in Measuring Survey Bias Using Wordscores

Here I offer some guidelines in implementing and interpreting a Wordscores model for text bias. Firstly, this method is not a replacement for rigorous survey pretesting; rather, it is a complement to it. In writing a survey, rather than

selecting the question my model estimates as having the least absolute bias, it is best used as a filter to select the *set* of least biased questions, then exploring those in a pretesting framework.

Secondly, it is extremely important to consider whether the supervision labels accurately reflect the underlying dimension of interest. In exploring the bias in survey questions, I hypothesize that the language Members of Congress use in their press releases and similar contexts induce subtle cues that constituents pick up on when reading survey questions. However, if the scale on which these cues exist is not well-captured by roll call votes on the floor of Congress as scaled by DW-NOMINATE, or if it is captured with significant measurement error, then the model may be noisy. While I have shown that in this application the model is robust to this concern, the model may require different training labels for estimating partisanship along a dimension other than the first dimension of DW-NOMINATE.

A third consideration is the size difference between the training documents and the test documents. Since a document's bias is an additive function of its words' bias, if the test set is systematically shorter than the training set, its estimated bias will be on average closer to 0 than the documents in the training set. This is not a concern when examining relative bias as I do in this paper; as well, Wordscores inherently rescales its predicted scores so their variance matches that of the training set.

A fourth and much more subtle problem relates to the temporal relationship between the training set and the test set. It is probable that the partisan connotations of words change dramatically over even very short periods of time. A survey question relating to healthcare estimated using a training set back to 2000 likely underestimates the bias for the words "affordable" and "care", as those words came into prominence as Democratic-biased after the introduction of the Affordable Care Act in 2011. The two ways to solve this problem are to carefully select a time scale for the training set, or with a large enough corpus, introduce interactions between specific words and time, parametrically allowing the effect of a word to change over time.

Finally, since it may be that the relationship between word use and partisanship varies by issue area, it is often useful to carefully select the training set by topics. For example, the word "free" may be a very liberal word when referenced in the context of "free healthcare" or "free education", it may be conservative in reference to "free trade" or "free market". While one can avoid these specific examples by including both unigrams and bigrams in the feature set (as I do in this paper), it is also advisable to use models trained on a corpus as similar as possible to the intended prediction set.

In summary, when using this method:

1. Gather a training corpus which accurately reflects the underlying dimension of interest and document labels which reliably measure that dimension for each document.[14]
2. Determine how much of the corpus is relevant, and down-weight or prune irrelevant training documents.[15]

[14] In this paper, I use DW-NOMINATE as the training labels, which are the gold standard for measuring Congressional ideology, and Congressional floor speech and press releases as the training corpus, which presumably capture the universe of partisan policy.

[15] I remove some procedural speech from the Congressional floor speech corpus since it is stylistically very distinct from policy-oriented language of survey questions.

3. Determine if your test set is time-sensitive, and down-weighting older observations or allow for time interactions.[16]
4. From the corpus, build an appropriately-sized feature set. Experimenting with different feature pruning thresholds is encouraged.[17]
5. Run a classification method of your choice according to your applications. Wordscores is computationally efficient and induces no sparsity, though classifier ensembles or neural nets may provide better predictions with sufficient training data.
6. Interpret (relative) bias scores with a cautious eye!

7 Conclusions & Further Research

I make two key contributions in this paper. First, I introduce a model for estimating the bias in survey questions at scale, and in a series of survey experiments, I show that it outperforms public opinion researchers and professionals in predicting which questions are more biased than others. I make available training data, code, and an online interface to allow other researchers and professionals to examine their own questions and others'. The web application, written in R Shiny, is hosted on GitHub[18].

Second, I explore a temporally fine-grained data set of survey questions from 1994 to 2017. I show that while questions related to the economy are stable in bias over time, questions related to Barack Obama become steadily more conservative over the course of his term. Questions related to abortion and national security are consistently more conservative than average; questions related to education and gun control are consistently more liberal than average. As well, most survey firms produce, on average, conservatively biased questions, casting serious doubt on the industry's capacity as a standard for objective measurement.

In ongoing work, I examine two themes arising from this paper. The first theme explores the underlying psychological mechanisms that produce framing effects based on word choice. If respondents reading a biased question are reacting to cues from elite rhetoric, it is likely that respondents with more exposure to political rhetoric or greater political knowledge will be more susceptible to framing effects. This runs counter to evidence showing that it is exactly the political neophytes who are most susceptible to survey design effects or other manipulations, but follows closely with work by Zaller (1992) arguing that political sophisticates behave most like their copartisan elites. Follow-up research explores the results of an experiment testing this hypothesis, and its implications for our understanding of how political sophistication interacts with survey bias.

The second set of ongoing work extends this model to consider forms of bias other than partisan bias, for example racial and gender bias. Hopefully, with widespread adoption, these tools can aid public opinion researchers in

[16] I provide robustness checks to this in Appendix D.
[17] I trim sparse features, but keep all common ones, in the interest of computational and memory efficiency.
[18] https://github.com/aaronrkaufman

constructing more valid estimates of aggregate preferences, to help journalists to more reliably reporting poll data, and to assist the public in becoming more informed consumers of survey results.

References

Achen, C. H. and Bartels, L. M. (2016). *Democracy for realists: Why elections do not produce responsive government*. Princeton University Press.

Albertson, B. L. (2015). Dog-whistle politics: Multivocal communication and religious appeals. *Political Behavior*, 37(1):3–26.

Ansolabehere, S. and Iyengar, S. (1994). Of horseshoes and horse races: Experimental studies of the impact of poll results on electoral behavior. *Political Communication*, 11(4):413–430.

Ansolabehere, S., Rodden, J., and Snyder, J. M. (2008). The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review*, 102(2):215–232.

Areni, C. S., Ferrell, M. E., and Wilcox, J. B. (2000). The persuasive impact of reported group opinions on individuals low vs. high in need for cognition: Rationalization vs. biased elaboration? *Psychology & Marketing*, 17(10):855–875.

Arlot, S., Celisse, A., et al. (2010). A Survey of Cross-Validation Procedures for Model Selection. *Statistics Surveys*, 4:40–79.

Armstrong, J. S. and Overton, T. S. (1977). Estimating nonresponse bias in mail surveys. *Journal of marketing research*, pages 396–402.

Baek, Y. M., Cappella, J. N., and Bindman, A. (2011). Automating content analysis of open-ended responses: Wordscores and affective intonation. *Communication methods and measures*, 5(4):275–296.

Bélanger, É. and Meguid, B. M. (2008). Issue salience, issue ownership, and issue-based vote choice. *Electoral Studies*, 27(3):477–491.

Benoit, K. and Nulty, P. (2016). quanteda: Quantitative analysis of textual data. *R package version 0.9*, 8.

Berinsky, A. J., Huber, G. A., and Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Political Analysis*, 20(3):351–368.

Boudreau, C. and McCubbins, M. D. (2010). The blind leading the blind: Who gets polling information and does it improve decisions? *The Journal of Politics*, 72(2):513–527.

Brown, P. F., deSouza, P. V., Mercer, R. L., Della Pietra, V. J., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4).

Butler, D. M. (2011). Can Learning Constituency Opinion Affect How Legislators Vote? Results from a Field Experiment. *Quarterly Journal of Political Science*, 6(1):55–83.

Campbell, A., Converse, P. E., Miller, W. E., and Stokes, D. E. (1960). *The American Voter*. Wiley.

Carroll, R., Lewis, J. B., Lo, J., and Rosenthal, H. (2009). Measuring Bias and Uncertainty in DW-NOMINATE. *Political Analysis*, 17(3).

Chong, D. and Druckman, J. N. (2007). Framing Public Opinion in Competitive Democracies. *American Political Science Review*, 101(04):637–655.

Converse, P. E. (1964). The nature of belief systems in mass publics. In Apter, D. E., editor, *Ideology and Discontent*. Free Press, New York.

Denny, M. J. and Spirling, A. (2018). Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis*, pages 1–22.

Edelman, M. (2013). *Political language: Words that succeed and policies that fail*. Elsevier.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., and Stewart, B. M. (2018). How to make causal inferences using texts.

Enos, R. D., Hill, M., and Strange, A. M. (2016). Voluntary digital laboratories for experimental social science: The harvard digital lab for the social sciences. *Working Paper*.

Entman, R. M. (2007). Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.

Fong, C. and Grimmer, J. (2016). Discovery of treatments from text corpora. In *ACL (1)*.

Gelman, A., Goel, S., Rivers, D., Rothschild, D., et al. (2016). The mythical swing voter. *Quarterly Journal of Political Science*, 11(1):103–130.

Goodman, J. K., Cryder, C. E., and Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, 26(3):213–224.

Grimm, P. (2010). Social desirability bias. *Wiley International Encyclopedia of Marketing*.

Guthrie, D., Allison, B., Liu, W., Guthrie, L., and Wilks, Y. (2006). A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, pages 1–4. sn.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A*, 171(2):481–502.

Iyengar, S. (1990). The accessibility bias in politics: Television news and public opinion. *International Journal of Public Opinion Research*, 2(1):1–15.

Kam, C. D. and Utych, S. M. (2011). Close elections and cognitive engagement. *The Journal of Politics*, 73(4):1251–1266.

Kaufman, A., King, G., and Komisarchik, M. (2017). How to measure legislative district compactness if you only know it when you see it.

Key, Jr., V. O. (1966). *The Responsible Electorate: Rationality in Presidential Voting, 1936-1960*. Belknap Press of Harvard Univ Press, Cambridge.

Klemmensen, R., Hobolt, S. B., and Hansen, M. E. (2007). Estimating policy positions using political texts: An evaluation of the wordscores approach. *Electoral Studies*, 26(4):746–755.

Klüver, H. (2009). Measuring interest group influence using quantitative text analysis. *European Union Politics*, 10(4):535–549.

Krosnick, J. A., Malhotra, N., and Mittal, U. (2014). Public Misunderstanding of Political Facts: How Question Wording Affected Estimates of Partisan Differences in Birtherism. *Public Opinion Quarterly*, 78(1).

Lakoff, G. (2010). *Moral politics: How liberals and conservatives think*. University of Chicago Press.

Lang, K. and Lang, G. E. (1984). The impact of polls on public opinion. *The Annals of the American Academy of Political and Social Science*, 472(1):129–142.

Lasswell, H. D. (1965). Language of politics.

Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2).

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*.

Lowe, W. (2008). Understanding wordscores. *Political Analysis*, 16(4):356–371.

Lowe, W., Benoit, K., Mikhaylov, S., and Laver, M. (2011). Scaling policy preferences from coded political texts. *Legislative studies quarterly*, 36(1):123–155.

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., and Campos, L. F. (2017). Worth weighting? how to think about and use sample weights in survey experiments. *Political Analysis*.

Morwitz, V. G. and Pluzinski, C. (1996). Do polls reflect opinions or do opinions reflect polls? the impact of political polling on voters' expectations, preferences, and behavior. *Journal of Consumer Research*, 23(1):53–67.

Petrocik, J. R. (1996). Issue ownership in presidential elections, with a 1980 case study. *American journal of political science*, pages 825–850.

Petrocik, J. R., Benoit, W. L., and Hansen, G. J. (2003). Issue ownership and presidential campaigning, 1952–2000. *Political Science Quarterly*, 118(4):599–626.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88(5):879.

Roberts, M. E., Stewart, B. M., Tingley, D., Airoldi, E. M., et al. (2013). The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Rothschild, D. and Malhotra, N. (2014). Are public opinion polls self-fulfilling prophecies? *Research & Politics*, 1(2):2053168014547667.

Roy, J., Singh, S. P., Fournier, P., and Andrew, B. (2015). An experimental analysis of the impact of campaign polls on electoral information seeking. *Electoral Studies*, 40:146–157.

Russell, A. (2017). Us senators on twitter: Asymmetric party rhetoric in 140 characters. *American Politics Research*, page 1532673X17715619.

Schattschneider, E. (1960). *The Semisovereign People*. Holt, Rinehart and Winston, New York.

Seeberg, H. B. (2017). How stable is political parties issue ownership? a cross-time, cross-national analysis. *Political Studies*, 65(2):475–492.

Sink, J. (2012). Poll: Plurality say polls biased for obama.

Skalaban, A. (1988). Do the polls affect elections? some 1980 evidence. *Political Behavior*, 10(2):136–150.

Sniderman, P. M. and Theriault, S. M. (2004). The structure of political argument and the logic of issue framing. In Saris, W., editor, *Studies in Public Opinion*, pages 133–65. Princeton University Press.

Sonck, N. and Loosveldt, G. (2010). Impact of poll results on personal opinions and perceptions of collective opinion. *International Journal of Public Opinion Research*, 22(2):230–255.

Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin*, 133(5):859.

Tversky, A. and Kahneman, D. (1985). The framing of decisions and the psychology of choice. In *Environmental Impact assessment, technology assessment, and risk analysis*, pages 107–129. Springer.

Utych, S. M. and Kam, C. D. (2013). Viability, information seeking, and vote choice. *The Journal of Politics*, 76(1):152–166.

Wodak, R. (1989). *Language, power and ideology: Studies in political discourse*, volume 7. John Benjamins Publishing.

Zaller, J. R. (1992). *The Nature and Origins of Mass Opinion*. Cambridge University Press, New York.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1):75–98.

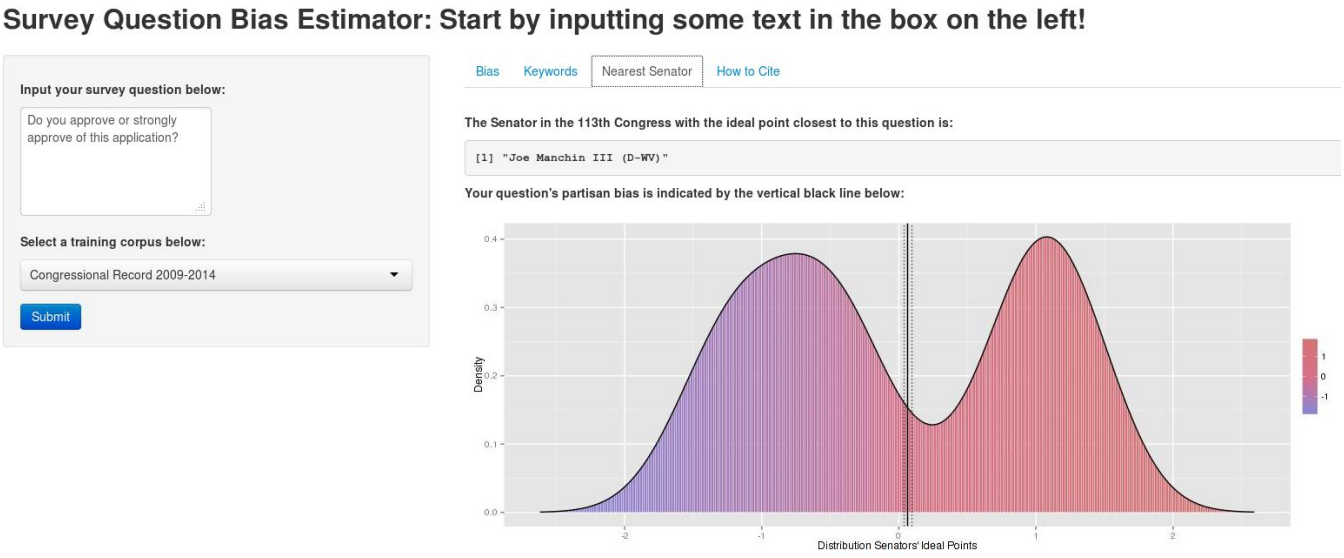
Zwane, A. P., Zinman, J., Van Dusen, E., Pariente, W., Null, C., Miguel, E., Kremer, M., Karlan, D. S., Hornbeck, R., Giné, X., et al. (2011). Being surveyed can change later behavior and related parameter estimates. *Proceedings of the National Academy of Sciences*, page 201000776.

Appendices

A Web Implementation of the Wordscores Model

While the importance of pretesting for internal validity cannot be understated, collecting a training corpus, converting a potential survey into term-document matrices, and running classifier ensembles may similarly time-intensive, computationally difficult, and monetarily expensive, especially for those survey researchers without programming expertise. Since the purpose of the method is to reduce time and monetary costs to pretesting surveys, I have developed an application built in R Shiny [19] to implement this method as it currently exists. Users input a survey question, or series of survey questions, and this application produces a set of information related to the text's partisan bias: the raw bias score, a confidence interval, the key lexical features which contribute to the score, the Member of Congress with an ideal point closest to the inputted text, and a density plot of members of the 115th Congress by ideal point with a vertical line indicating the survey question's bias. For images of the application, see Figure 8. The application is currently available at `http://hmdc.shinyapps.io/SurveyBias`.

Figure 8: A screenshot of the R Shiny implementation of my Wordscores model. Users input a survey question on the left, and optionally identify a topic-specific model to use. The text's partisan bias is estimated, and plotted in relation to Senators of the 113th Congress. On additional tabs are breakdowns of the most partisan n-grams in the text document and instructions for interpreting results.



[19] `http://shiny.rstudio.com/`

B Additional External Validity Results

I include two related questions in the results from Section 4. I randomly assigned one of four questions about abortion, and an additional randomized choice about firearms alongside the three in Table 2. In both of these studies, my model outperforms human experts.

Question Text	Revealed Rank	Model Rank	Expert Rank
Abortion should be illegal in all cases, including in cases of rape and incest.	1	**1**	4
Abortion should be illegal in all cases, except in cases of rape and incest.	2	**2**	**2**
Abortion should be illegal in all cases.	4	**3**	**5**
Abortion should be illegal, except with parental permission.	3	**4**	1
Abortion should be illegal, except when the life of the mother is threatened.	5	**5**	3
We should require mental health checks before anyone can purchase a firearm.	1	**1**	2
We should require background checks before anyone can purchase a firearm.	2	**2**	1
Even considering the importance of protecting oneself, we should require background checks before anyone can purchase a firearm.	3	**3**	**3**
Even considering the importance of the Second Amendment protecting the right to bear arms, we should require background checks before anyone can purchase a firearm.	4	**4**	**4**

Table 3: The Wordscores model predicts the relative framing bias generally better than the aggregated expert ranks. For each question, the closer rank is bolded.

C Wordscores

Wordscores, developed by Laver, Benoit, and Garry in 2003 and used across a wide span of political science applications (Klemmensen et al., 2007; Klüver, 2009; Baek et al., 2011; Lowe et al., 2011), is an inductive method designed for extracting policy positions from text by examining similarities and dissimilarities of word usage across documents. In the authors' own words, they "... use the relative frequencies we observe for each of the different words in each of the reference texts to calculate the probability that we are reading a particular reference text, given that we are reading a particular word" (page 313).

For example, if we know that Democratic texts use the word "undocumented" 30 times per 10,000 words and Republicans use that word only 10 times per 10,000 words, then upon observing the word "undocumented" in a text, our best guess is that there is a 75% chance our document comes from a Democrat. If we assume the average DW-NOMINATE score of a Democrat is -1 and the average score of a Republican is +1, then a document consisting only of the word "undocumented" has a score of $0.75 \times (-1) + 0.25 \times (+1) = -0.5$.

Notation

Drawing heavily from Laver, Benoit, and Garry (2003), pages 316-217, I explain in detail the math underlying Wordscores.

For a set of training documents T with training labels Y_t, and a vocabulary of n-grams w, the frequency of word w in document t as a proportion of the entire document is $F_{w,t}$. Therefore the probability that we are reading document t given that we see word w is:

$$P_{w,T=t} = \frac{F_{w,T=t}}{\sum_t F_{w,t}} \tag{4}$$

We can then calculate the score of any word along the dimension of Y_t by summing across documents:

$$S_w \sum_r P_{w,T=t \times Y_t} \tag{5}$$

Having compiled the scores for all words in the vocabulary, calculating the total score for a new document v is trivial:

$$Y_v = \sum_w F_{v,w} \times S_w \tag{6}$$

Finally, to compensate for the underdispersion of scores in new test documents, we rescale each test document's score to Y_{v*}. In the case of n test documents, test docuemnt v's rescaled score is calculated as follows:

$$Y_{v*} = (Y_v - \frac{1}{n}\sum_v Y_v)(\frac{SD_t}{SD_v}) + \frac{1}{n}\sum_v Y_v \tag{7}$$

where SD_t and SD_v measure the dispersion of the training and test set scores, respectively.

Drawbacks & Assumptions of a Wordscores model

I distinguish between assumptions and drawbacks in my model which derive from Wordscores, and those deriving from the "bag of n-grams" assumption. The primary assumption in Wordscores, as its authors write on page 314, is that the training labels are confident and accurate; incidentally, this is an assumption in any modeling approach.

Wordscores may seem to make assumptions such as: a document's bias is the sum of its individual words' biases. However, this need not be the case, as Wordscores can easily include interaction terms or skip n-grams (Guthrie et al., 2006) as features. It also appears to assume that negation ("We are NOT pro-choice, pro-immigrant, pro-gun control") does not matter, but that is a feature of the "bag of n-grams" assumption rather than Wordscores, and regardless, Wordscores will provide robust results so long as such negated statements are not overly represented in the corpus, in which case including skip n-grams as features is highly desirable.

The most important assumption in Wordscores is that the training set documents and test set documents are substantively similar. This means that the true probability of a document being a liberal document given word p in the training set T is equal to that quantity in the test set V:

$$P(Y|w,i) = P(Y|w,j) \forall i \in T, j \in V \tag{8}$$

This assumption is difficult to verify, but the results in Section 4 suggest that it holds sufficiently well in practice.

D Temporal Variation in Word Valence

In Section 5 I argue that contrary to a common media narrative, surveys were biased *against* Barack Obama during his presidency. Examining trends by six major survey firms, I show that each asked increasingly conservative questions relative to the average survey about him prior to his inauguration.

However, we might suspect that the strong conservative trend in survey questions related to Barack Obama is a statistical artifact related to either the changing composition of Congress or the changing use of words. For example, more Republicans entered Congress in 2010 and 2014, common words with little to no partisan valence such as prepositions are used more by Republicans than Democrats; similarly, questions related to African American affairs fielded in the 1940s would be unpalatable today for some of the language used. However, this cannot affect these results for two reasons. Since my model is trained on the *entire corpus* at once rather than on individual years, the effect of changing Congressional composition is a mean shift in individual word scores. And since I examine trends relative to a 2007 baseline, all bias scores are de-meaned *ex ante*.

We might also attribute these trends to compositional changes. Imagine there are only two questions about Obama – a liberal question, and a conservative question. We would observe the same trends if the conservative question is asked more frequently in later years than in earlier years. We do observe that there are fewer liberal-biased questions related to Obama after 2012 than before; those questions, and the liberal-biased questions in 2012, are largely about Presidential elections. The increase in conservative bias in 2016 comes from an increase in questions using the term "Obamacare," which my model estimates has a conservative bias of +0.52 ("Obama" is mostly neutral at +0.02). Using a topic model to examine the changes in question composition over time, however, shows that there are no general compositional trends in question topics that explain the increasingly conservative bias among many of the major news outlets.

As a test of the variability among word scores during Obama's first term, I calculate four separate Wordscores models, specific to each of his first four years. I present the coefficients related to Obama below. If it is the case that words related to Obama are becoming more conservative over time, that might explain the trend. While five of the six terms become more conservative, and the average change in those six terms is 0.05 , that might explain less than one fifth of the approximately 0.25 increase in conservative bias.

Table 4: The word scores for features related to Barack Obama from 2009 to 2012.

word	2009	2010	2011	2012
obama	0.01	0.13	0.17	0.18
obamasaid	0.22	0.23	0.13	0.25
presidobama	0.06	0.08	0.14	0.13
obamacar	0.71	0.58	0.59	0.42
barackobama	-0.03	0.16	0.18	0.15
obamaadministr	0.07	0.17	0.22	0.23