Environmental Data Science (2024), 3: e10, 1–17
doi:10.1017/eds.2023.44

APPLICATION PAPER

Ideology from topic mixture statistics: inference method and example application to carbon tax public opinion

Maximilian Puelma Touzel[1,2] and Erick Lachapelle[3]

[1]Mila, Québec AI Institute, Montréal, QC, Canada
[2]Department of Computer Science and Operations Research, Université de Montréal, Montréal, QC, Canada
[3]Department of Political Science, Université de Montréal, Montréal, QC, Canada
Corresponding author: Maximilian Puelma Touzel; Email: max.puelmatouzel@gmail.com

Received: 02 February 2023; **Revised:** 24 August 2023; **Accepted:** 24 October 2023

Keywords: artificial intelligence; carbon tax; climate change policy; data science; machine learning; public opinion; topic model

Abstract

Political opposition to fiscal climate policy, such as a carbon tax, typically appeals to fiscal conservative ideology. Here, we ask to what extent public opposition to the carbon tax in Canada is, in fact, ideological in origin. As an object of study, ideology is a latent belief structure over a set of issue topics—and in particular their relationships—as revealed through stated opinions. Ideology is thus amenable to a generative modeling approach within the text-as-data paradigm. We use the Structural Topic Model, which generates word content from a set of latent topics and mixture weights placed on them. We fit the model to open-ended survey responses of Canadians elaborating on their support of or opposition to a carbon tax, then use it to infer the set of mixture weights used by each response. We demonstrate this set, moreso than the observed word use, serves efficient discrimination of opposition from support, with near-perfect accuracy on held-out data. We then operationalize ideology as the empirical distribution of inferred topic mixture weights. We propose and use an evaluation of ideology-driven beliefs based on four statistics of this distribution capturing the specificity, variability, expressivity, and alignment of the underlying ideology. We find that the ideology behind responses from respondents who opposed the carbon tax is more specific and aligned, much less expressive, and of similar variability as compared with those who support the tax. We discuss the implications of our results for climate policy and of broad application of our approach in social science.

Impact Statement

What people think about climate change and climate policy strongly depends on their politics. To better understand the role played by ideology in shaping people's attitudes, we assess people's ideological thinking by analyzing the ways in which they explain their stance in open-ended survey responses. We develop new ways of measuring how topics are talked about together. Independent of what respondents discussed, we find people elaborating on opposition to the carbon tax did so in a more focused and structured way than did those who support a carbon tax. This suggests public opposition to the tax could be more easily galvanized politically, while also suggesting the design of more effective climate policy and communications.

This research article was awarded an Open Materials badge for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

1. Introduction

Personal values and prior beliefs can have a greater effect than immediate reward and direct evidence in shaping how individuals interpret the world and how they respond when prompted to opine about specific aspects of it. One helpful construct for understanding this cognitive and collective process is ideology. As a system of interconnected beliefs, ideology is invoked to explain the role of political partisanship and the causes of political polarization (Baldassarri and Gelman, 2008). However, how to analyze ideology as a correlation structure on beliefs is a methodological challenge yet unresolved (Kalmoe, 2020). In this study, we propose an approach to infer properties of a latent ideology from open-ended survey responses, and ask to what extent these properties reveal differences in the way opposing groups rationalize their positions on carbon tax policy. Given the well-documented anti-tax rhetoric of conservative political discourse (Lakoff, 2010), we expected to observe differences between supporters and opponents of carbon tax policy.

There are many societal challenges for which methods that reveal ideology would be useful. A prime example is mitigating the worst effects of climate change. Climate change mitigation is in many countries a hotly contested political issue that is characterized by extensive ideological and partisan conflict (McCright and Dunlap, 2011; McCright et al., 2016; Birch, 2020). Addressing this challenge requires significant government intervention to shift energy production and consumption away from CO_2-emitting fossil fuels (McGlade and Ekins, 2015; Welsby et al., 2021). In this context, a price on carbon emissions is often proposed by economists as being simple, flexible, and easy to put on a rising schedule, encouraging emissions reductions across economic sectors at a low aggregate cost (Baranzini et al., 2000; Aldy and Stavins, 2012). Policy design has addressed public resistance to carbon tax policy measures via the targeted use of tax revenue in the form of lump sum rebates to taxpayers (Klenert et al., 2018). Despite these mechanisms, however, a lack of public acceptability remains a major obstacle to the adoption of substantive carbon taxes across countries (Lachapelle, 2017; Haites, 2018; Rabe, 2018).

While mediated by various factors (Davidovic et al., 2020), public support for carbon pricing is typically driven by ideology, with conservatives on the right of the political spectrum generally in opposition, and those on the left generally supporting the policy (Drews and Van, 2016). This structuring of carbon tax policy attitudes by ideology is evident even in countries that have implemented tax-and-rebate schemes designed to build public support by providing rational, monetary incentives that offset the policy's costs to consumers. For instance, at current pricing levels in Canada (one of few countries that have implemented a tax-and-rebate scheme), 8 of 10 households are estimated to receive more than they pay as a result of the policy, and yet roughly 8 out of 10 surveyed conservatives oppose the policy (Mildenberger et al., 2022). Mildenberger et al. (2022) probe that result using interventions and find conservatives underestimate the size of the rebate they receive. Such behavior is not necessarily irrational: strong priors can make rational decision-making insensitive to new data (Gershman, 2019). Even after being provided with information on the size of their rebate, individuals may continue to believe that they are net losers. One explanation for these results is that a broader system of values—an ideology—underlies subjects' decision-making. This putative belief system would then also likely manifest when subjects justify their opposition to the tax. A better understanding of how ideology underlies carbon tax opinion would explain the effectiveness of opposition campaigns centered around it and could also inform the design of more effective carbon pricing policy in general, perhaps via more effective communication of the policy's benefits. This might involve first identifying, then appealing to (and in some instances directly challenging) issue frames commonly associated with carbon tax policy beliefs.

Issue frames, when written out, can be partially quantified by the semantic content of the writing. Quantitative semantic representations typically focus on the frequency of word-use.[1] However, single-word frequency measures do not expose how the same words can be used when talking about different things. Other widely used approaches, such as sentiment analysis, classify responses into only a few

[1] This includes more refined frequency measures such as Tfidf (see Figure 3).

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

affective classes (“like”/“dislike”). By formulating a rich latent topic structure, *topic models* address both of these limitations. Topic models are now an established approach to understanding human-generated responses in the social sciences (Grimmer et al., 2022). The Structural Topic Model in particular has been applied to understand open-ended responses on a carbon tax in Spain (Savin et al., 2020), Norway (Tvinnereim et al., 2017), and the US (Povitkina et al., 2021). We were motivated to make a similar application to data collected in Canada. Unlike these previous works, however, we chose to focus on the correlated statistics of the weights placed on different topics as a means to interrogate ideology.

Our contribution is a set of measures and analyses that characterize latent topic mixture structure as a means of inferring ideology from open-ended survey responses. We also contribute the results of applying our approach to carbon tax opinion in Canada. To further motivate this application, we hypothesized that opposition to carbon taxes arises from a well-worn “tax is bad” ideology, involving a handful of correlated ideas (e.g., “distrust in government,” “unfairness,” “infringement on personal freedom”) that mutually enforce each other (Lakoff, 2010). Here, we use the Canadian data set collected by (Mildenberger et al., 2022), unique in the richness of its meta-data, to fit the parameters of a generative bag-of-words model of word responses having a latent structure of topics. We fit models to different response types: the subset of responses of those who stated separately that they support the tax, of those who stated they opposed the tax and of those two groups combined. After validating the fitted models, we use it to infer topic structure conditioned on the support and oppose response groups. We focus on the ways in which respondents mix topics when supporting or opposing the carbon tax. We not only find that responses are highly discriminable using these topic mixtures, but that there are clear differences in the topic mixture structure of the two response types that support our hypothesis.

2. Methods

2.1. Dataset

We analyzed a dataset of 3313 open-ended survey responses from respondents living in Canada’s four largest provinces (Ontario, Quebec, Alberta, and British Columbia) as published in Mildenberger et al. (2022). In addition to asking respondents for a categorical opinion response in *support/opposition/not sure* (response type), demographic survey data were collected such as age, gender, and province. Lifestyle data (car use, residence environment), and belief data (partisanship [measured using party voting] and political ideology) were also collected. Ideology was measured with a unidimensional scale asking respondents, “In politics, people sometimes talk of left and right. Where would you place yourself on the following scale?” Response options ranged from 0 to 10 with extreme values labeled “Far to the left” (0) and “Far to the right” (10). The mid-point was also labeled as “Centre” (5). In line with research documenting a general trend in partisan dealignment (Dalton and Wattenberg, 2002), partisanship was measured with a vote choice question asking respondents, “If a federal election were held today, for which party would you vote for?” Reflecting the multiparty nature of Canadian politics, response options included items for each of Canada’s main federal political parties: Liberal, Conservative, New Democratic Party, People’s Party, Greens, and Bloc Québécois (for respondents in Quebec). An option was also included for “I would not vote.” In the analysis, we recorded Ideology and Partisanship into three categories each. See Mildenberger et al. (2022) for more details on this rich meta-data and see Supplementary Material for the specific reductions of categories that we used. In addition to categorical responses, a central open-ended question in the survey asked respondents to elaborate on why they chose the categorical response *support/oppose/not sure.* For this analysis, we preprocessed these open-ended responses. French responses were first translated into English using the Google Translate API. The response corpus was pre-processed through automated spell-checking, removal of stop words, and reduction of word stems. Words were tokenized: each word in the vocabulary was assigned an index and each response was transformed into a vector representation of word counts in this vocabulary.

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

2.2. Word relevance

We used a standard term-relevance measure that combines term frequency (as a measure of term relevance) and inverse document frequency (as a measure of term specificity):

- *term frequency*, $\mathrm{tf}(v,d)$. The frequency of a vocabulary word v in a document d is $\mathrm{tf}(v,d) := \frac{n(v,d)}{n(d)}$, where $n(v,d)$ is the number of times the term v is in the document d and $n(d) = \sum_{v \in \mathcal{V}} n(v,d)$ is the total number of words in the document d. The term frequency over all the documents is then,

$$\mathrm{tf}(v) := \frac{\sum_{d=1}^{D} n(d)\mathrm{tf}(v,d)}{N_{\mathrm{total}}},$$

  where the denominator $N_{\mathrm{total}} = \sum_{d=1}^{D} n(d)$ is just the total word count across all D considered documents.
- *inverse document frequency*, $\mathrm{idf}(v)$. Here,

$$\mathrm{idf}(v) = \frac{\log(D+1)}{\log(n(v)+1)+1},$$

  where $n(v)$ is the number of documents in which the term v appears, that is, $n(v,d) > 0$. Idf is like an inverse document frequency.

Term frequency-inverse document frequency (Tfidf) for a word-document pair is then defined simply as the product, $\mathrm{Tfidf}(v,d) := \mathrm{tf}(v,d)\mathrm{idf}(v)$. This is then averaged over documents to create a word-only measure. We used the *sklearn* package implementation of Tfidf that averages the normalized values to remove dependence on the length of a document

$$\mathrm{Tfidf}(v) = \frac{1}{D}\sum_{d=1}^{D} \frac{\mathrm{Tfidf}(v,d)}{\|\mathrm{Tfidf}(\cdot,d)\|},$$

where $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{N} x_i^2}$ is the Euclidean norm (throughout the paper, we denote vectors with boldface symbols). As a word relevance metric, Tfidf adds discriminability than using frequency alone by downweighting words that appear in many documents, since these common words are less discriminative. We computed word clouds and rank histograms using the Tfidf values.

2.3. Response-type classification

We aimed to predict the response type label (*oppose*/*support*) for each response, first using the word-level representation (i.e., which words appear and how many times), then using the inferred topic mixture weights obtained from specific (e.g., K-dependent) fitted STMs on the combined dataset of support and oppose responses (see the following sections for details on the topic model). As preprocessing, we applied maximum absolute value rescaling to handle sparsity in the data for the word-level representation, and standard rescaling in the topic mixture case. In both cases, we then performed logistic regression. We then ranked features (words or topic weight vector components, respectively) by their weighted effect size and then performed logistic regression trained a classifier for each subset of ranked features of increasing number, n. For each n, we ran 100 trials of a 2:1 train/test split and report the mean and variance of classification accuracy on the test set.

2.4. The Structural Topic Model

Topic models are generative models that generate a set of words in a response document according to some distribution. Topic models are typically *bag of word* models, which eschew grammar and syntax to focus only on the content and prevalence of words (i.e., sampling distributions do not condition on previously sampled words in a response). We select a topic model with rich latent structure: the *Structural Topic Model* (STM) (Roberts et al., 2014). Like the Correlated Topic Model (Blei and

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

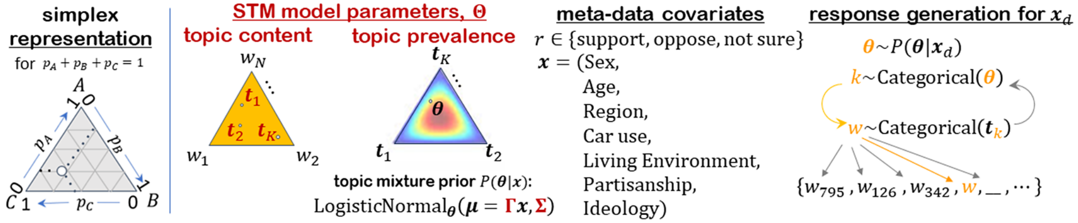
simplex representation
for $p_A + p_B + p_C = 1$
A
B
C
1
0
1
0
1
0
p_A
p_B
p_C
STM model parameters, Θ
topic content
topic prevalence
w_N
t_1
t_2
t_K
w_1
w_2
t_K
$\boldsymbol{\theta}$
t_1
t_2
topic mixture prior $P(\boldsymbol{\theta}|\boldsymbol{x})$:
$\text{LogisticNormal}_{\boldsymbol{\theta}}(\boldsymbol{\mu} = \boldsymbol{\Gamma x}, \boldsymbol{\Sigma})$
meta-data covariates
$r \in \{\text{support, oppose, not sure}\}$
$\boldsymbol{x}$ = (Sex,
Age,
Region,
Car use,
Living Environment,
Partisanship,
Ideology)
response generation for $\boldsymbol{x}_d$
$\boldsymbol{\theta} \sim P(\boldsymbol{\theta}|\boldsymbol{x}_d)$
$k \sim \text{Categorical}(\boldsymbol{\theta})$
$w \sim \text{Categorical}(\boldsymbol{t}_k)$
$\{w_{795}, w_{126}, w_{342}, w, \_, \cdots\}$

Figure 1. *The Structural Topic Model for analysis of public opinion survey data. From left to right: Recall the simplex as the space of frequency variables. STM parameters define topic content through the topics' word frequencies and topic prevalence through parameters associated with the topic mixture prior. We use seven meta-data covariates. We also store the respondent's categorical response (in support, in opposition, or not sure) to the issue question (here carbon tax). By sequentially sampling topics and words, the model produces a bag-of-words open-ended response elaborating on this opinion.*

Lafferty, 2005), the STM uses a logistic normal distribution from which to sample the topic mixture weights on a document and can thereby exhibit arbitrary topic-topic covariance via the covariance parameter matrix of the logistic normal distribution. Unlike the Correlated Topic Model in which the mean and covariance matrix of the logistic normal distribution are learned parameters, the STM allows for parametrizing these using a linear model of the respondents' meta-data. We discuss our choice of meta-data model below. In the standard implementation of STM that we use here (Figure 1), only the mean of the logistic normal is made dependent on the meta-data. This adds flexibility beyond that offered by the CTM and makes the STM appropriate for datasets with non-trivial latent topic structure, as we expect here. Specifically, our use of the STM model specifies the parameter tuple $\boldsymbol{\Theta} = \left(\{\boldsymbol{t}_k\}_{k=1}^K, \Sigma, \Gamma\right)$ as follows (Figure 1):

- *Topic content*: an underlying set of K topics indexed by k, $\{\boldsymbol{t}_k\}_{k=1}^K$, where each topic, $\boldsymbol{t}_k = \left(\beta_{k,1}, \ldots, \beta_{k,|\mathcal{V}|}\right)$, is a list of word frequencies on a given vocabulary $\mathcal{V}$ indexed by v ($\sum_{v=1}^{|\mathcal{V}|} \beta_{k,v} = 1$ for all k), and
- *Topic prevalence*: a prior distribution $P(\boldsymbol{\theta}|\boldsymbol{x})$ on the topic mixture vector, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ (weights are normalized, $\sum_{k=1}^K \theta_k = 1$) that is conditioned on the meta-data class, $\boldsymbol{x}$. Here, the distribution is chosen as a Logistic Normal with covariance matrix parameter Σ assumed to be the same across meta-data classes, and mean $\boldsymbol{\mu} = \Gamma \boldsymbol{x}$ is a linear transformation of $\boldsymbol{x}$ with transformation matrix parameter Γ.

A single response sample for a given response document of length N words that comes from a respondent in meta-data class $\boldsymbol{x}$ is generated by sampling a topic mixture weight vector $\boldsymbol{\theta}$ once from the weight vector distribution $P(\boldsymbol{\theta}|\boldsymbol{x})$, then iteratively sampling a topic index, $z_n \in \{1, \ldots, K\}$, from $\text{Categorical}_K(\boldsymbol{\theta})$ and a word, $w_n \in \mathcal{V}$ from the distribution formed from that topic's frequencies, $\text{Categorical}_{|\mathcal{V}|}(\boldsymbol{t}_{k=z_n})$, N times to make the response $\{w_1, \ldots, w_N\}$, with $n = 1, \ldots, N$. Topic frequencies are represented in log space according to the Sparse Additive Generative text model (Eisenstein et al., 2011). This topic content model allows for meta-data dependence. However, again confronted with the challenging model selection problem, and without strong hypotheses about how demographic-biased word use affects topic correlations (our primary interest), we chose to leave out this dependence.

2.5. *STM model fitting*

We built a Python interface for a well-established and computationally efficient STM inference suite written in the R programming language (Roberts et al., 2019). This STM package has a highly optimized

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

parameter learning approach that largely overcomes the inefficiency of not using a conjugate prior.[2] In particular, STM uses variational expectation maximization, with the expectation made tractable through a Laplace approximation. Accuracy and performance of the method are improved by integrating out the word-level topic and through a spectral initialization procedure. For details, see (Roberts et al., 2014). The Python code that we developed to access this package is publicly accessible with the rest of the code we used for this paper in the associated GitHub repository.

The parameter fitting performed by this STM package uses a prior on its covariance matrix parameter Σ of the topic mixture weight prior that specifies a prior on topic correlations: $\sigma \in [0,1]$ giving a uniform prior for $\sigma = 0$ and an independence (.e. diagonal) prior for $\sigma = 1$ (see the package documentation for more details on σ). Unless otherwise stated, we used the default value of $\sigma = 0.6$. Note that our main analysis of statistical measures does not use Σ. The package also offers priors on Γ. We used the recommended default "Pooled" option. This uses a zero-center normal prior for each element, with a Half-Cauchy(1,1) prior on its variance.

We fit three model types based on what dataset they are trained on. Here, a dataset $\mathcal{D} = \{(r_d, \boldsymbol{w}_d, \boldsymbol{x}_d)\}_{d=1}^{D}$ is a set of (categorical response, open-ended response, and respondent metadata) tuples. We fit one model to each of the support and oppose respondent data subsets, $\mathcal{D}_{\text{support}}$ and $\mathcal{D}_{\text{oppose}}$, as well as one to a combined dataset $\mathcal{D}_{\text{combined}} = \mathcal{D}_{\text{support}} \bigcup \mathcal{D}_{\text{oppose}}$ that joins these two subsets of tuples into one. The model parameters obtained from maximum likelihood fitting to a dataset are denoted $\widehat{\boldsymbol{\Theta}}$. As with other topic models, an STM takes the number of topics, K, as a fixed parameter (on which most quantities depend so we omit it to simplify notation). We thus obtained three K-dependent families of model parameter settings $\widehat{\boldsymbol{\Theta}}_m$, for $m \in \{\text{support}, \text{oppose}, \text{combined}\}$. With model parameters in hand, we obtained our object of primary interest: the models' topic mixture posterior distributions conditioned on response type $r' \in \{\text{support}, \text{oppose}\}$,

$$P\left(\boldsymbol{\theta} | \boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}_m, r'\right) = \sum_{\mathbf{x}} P_m(\boldsymbol{\theta}|\boldsymbol{x}) p(\boldsymbol{x}|r') \tag{1}$$

$$\approx \frac{1}{|\mathcal{D}_{r'}|} \sum_{d \in \mathcal{D}_{r'}} \text{LogisticNormal}(\Gamma_m \boldsymbol{x}_d, \Sigma_m), \tag{2}$$

where we have substituted the Logistic Normal distribution for $P_m(\boldsymbol{\theta}|\boldsymbol{x})$ and use the empirical average over $\boldsymbol{x}$. To compare the ideologies behind support and oppose responses, we make the two comparisons from evaluating equation (1):

$$P\left(\boldsymbol{\theta}|\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}_{\text{support}}, \text{support}\right) \overset{?}{\leftrightarrow} P\left(\boldsymbol{\theta}|\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}_{\text{oppose}}, \text{oppose}\right) \text{ and} \tag{3}$$

$$P\left(\boldsymbol{\theta}|\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}_{\text{combined}}, \text{support}\right) \overset{?}{\leftrightarrow} P\left(\boldsymbol{\theta}|\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}_{\text{combined}}, \text{oppose}\right). \tag{4}$$

Since we are interested in topic correlations, we do not need to derive effect sizes for particular meta-data covariates, which in general depend on the particular meta-data model, here given by which elements of Γ in $\mu = \Gamma \boldsymbol{x}$ are non-zero and their sign. Consequently, we can sidestep the challenging meta-data model selection problem (Wysocki et al., 2022) by simply including all the measured covariates that might contribute (i.e., all elements of Γ are specified as real-valued free parameters). We thus selected a broad subset of the meta-data from Mildenberger et al. (2022): $\boldsymbol{x} = (\text{age}, \text{sex}, \text{region}, \text{car use}, \text{partisanship}, \text{ideology})$.

We analyzed results across a range of K to ensure validity of our conclusions. Models with larger K typically give higher model likelihood, but are at higher risk of over-fitting. The model likelihood on a heldout test dataset will however exhibit a maximum likelihood at finite K as overfitting to the remaining "train" samples becomes significant with increasing K. The value of K where the heldout likelihood achieves its maximum will depend on the relative size of the test data relative to the train data in the train-test

[2] The widely used Latent Dirichlet Allocation topic model uses a Dirichlet prior and the variational Bayes algorithm for this reason, but lacks expressiveness as a result.

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

split. Nevertheless, a loose upper-bound on K is the value at which the maximum of the split realization-averaged held-out likelihood for a train-biased train-test data split (here 9:1) is obtained. For the support, oppose, and combined response type models, maximums were observed at $K = 15, 18$, and 20, respectively for averages over 100 and 200 train-test split realizations for separate and combined response type models, respectively (see Figure 6e,j). Note that the strong subadditivity (20/(15+18) = 0.61) in these data-selected topic numbers suggests a high amount of overlap in the topic content between support and oppose responses.

We can directly compare the response-type conditioned posteriors of the combined model at any given K. In contrast, comparison of posteriors for the response-type specific models should account for the fact that the likelihood will in general suggest distinct K for each. With a prior in hand, we could simply marginalize over K in the posterior for each. We do not have such a posterior in hand, unfortunately. A natural alternative is simply to compare models at the K at which each achieves the respective maximum in the likelihood. We highlight these maximum comparisons when plotting results over K.

2.6. *Topic quality*

We assessed topic quality across different topic numbers using two related measures (the standard analysis in STM literature). First, exclusivity (high when a topic's frequent words are exclusive to that topic) was measured using the FREX score: the linearly-weighted harmonic mean of normalized rank of a word's frequency relative to other words within a topic ($\beta_{k,v}$) and the same across topics,

$$\mathrm{FREX}_{k,v} = \left(\frac{\omega}{\mathrm{NormRank}\left(\beta_{\cdot,v}\right)_k} + \frac{1-\omega}{\mathrm{NormRank}\left(\beta_{k,\cdot}\right)_v} \right)^{-1} \tag{5}$$

$$\mathrm{FREX}_k = \sum_{i=1}^{N_{\mathrm{top}}} \mathrm{FREX}_{k,v_i}(\omega), \tag{6}$$

where $\mathrm{NormRank}(\boldsymbol{x})$ returns the n-dimensional vector of n-normalized ascending ranks of the n-dimensional vector $\boldsymbol{x}$ (high rank implies high relative value) and ω is the linear weight parameter (set by default to 0.7). In this and the following definition, the notation v_i is the index having descending rank i in $\beta_{k,\cdot}$ (low rank implies high relative value, i.e., the more weighted topic words) and the sum is over the top N_{top} such words for topic k (the default setting of $N_{\mathrm{top}} = 10$ was used). Note that this implies that the second term simplifies using the identity $\mathrm{NormRank}\left(\beta_{k,\cdot}\right)_{v_i} = 1 - \frac{i}{|\mathcal{V}|}$. Second, semantic coherence (high when a topic's frequent words co-occur often), was defined as

$$C_k = \sum_{i=2}^{N_{\mathrm{top}}} \sum_{j=1}^{i-1} \ln\left(\frac{D\left(v_i, v_j\right) + 1}{D\left(v_j\right)} \right), \tag{7}$$

where $D(v)$ counts the documents (i.e., responses) in which the word v occurs and $D(v, v')$ counts the documents in which both words v and v' occur together. See the R package's documentation for more details (Roberts et al., 2019). We can study topic quality by analyzing the set of (FREX_k, C_k) pairs over topics in the plane spanned by exclusivity and coherence, where higher quality topics are positioned further to the top right. The mean of these points over topic for each value of k is used to evaluate particular STM models in Figure 4.

2.7. *Statistical measures of topic mixture statistics*

Our primary interest is in the statistical structure of the posteriors over topic mixture weight vectors, $\boldsymbol{\theta}$, equation (1). However, the geometry of a domain can have a strong effect on the estimation of statistical properties and estimation using the $\boldsymbol{\theta}$-representation is biased because of the normalization constraint on weight vectors. A now well-established approach to performing statistical analysis on variables defined on the simplex (Aitchison, 1982; Pawlowsky-Glahn and Egozcue, 2001) maps these variables to the space of logarithms of relative weights (relative to the geometric mean, $g(\theta) = \prod_{i=1}^{K} \theta_i^{1/K}$). This bijective transformation maps the simplex of K variables to $\mathbb{R}^{K-1}$ such that LogisticNormal distributions map

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

Table 1. *Four low-order statistics of a data cloud in a $(K-1)$-dimensional space ($\boldsymbol{\mu} \in \mathbb{R}^{K-1}$) with mean position $\overline{\boldsymbol{\mu}}$ and covariance matrix $\Sigma_{\boldsymbol{\mu}}$, having eigenvalues $\{\lambda_i\}_{i=1}^{K-1}$*

Data cloud property	Topic mixture interpretation	Normalized estimator
Position 	Specificity	$\frac{1}{\sqrt{K-1}}\mathbb{E}[\lVert\boldsymbol{\mu}\rVert]$, average distance from the origin (here: equal usage)
Size 	Variability	$\frac{1}{K-1}\sqrt{\sum_{i=1}^{K-1}\lambda_i}, \sqrt{\text{total variance}}$
Eccentricity 	(Lack of) Expressivity	$\frac{1}{K-1}\left(\sum_{i=1}^{K-1}\lambda_i\right)^2 / \left(\sum_{i=1}^{K-1}\lambda_i^2\right)$, intrinsic dimension
(+ve) Orientedness 	Alignment	$\frac{2}{(K-1)(K-2)}\sum_{\rho_{ij}>0, i<j}\rho_{ij}$, sum of positive correlation, $\left(\rho_{ij} = \frac{\left(\Sigma_{\boldsymbol{\mu}}\right)_{ij}}{\sigma_i\sigma_j}\right)$

back into their Normal counterparts. Applying this transformation to the computed models, we find that the average distance between all pairs from the set of estimated means $\widehat{\boldsymbol{\mu}}_d = \widehat{\Gamma}\boldsymbol{x}_d$ over the three datasets is typically much larger than the size of the variation arising from the fitted covariance matrix parameter $\widehat{\Sigma}$ (more than 95% of pairwise distances were larger than the total variance/dimension for three of the four comparisons (equations (3) and (4)) with the remainder at 85%; see Supplementary Material for details). Thus, the global geometry of the distribution is well-captured by the distribution of these means. We then focus on the empirical distribution of over $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{K-1})$, that is, the set $\{\widehat{\boldsymbol{\mu}}_d\}_{d=1}^{D}$, one for each response in the response type dataset and clustering according to demographic statistics.

We developed and used four intuitive characteristics of a data cloud (see Table 1). The first is simply the mean position $\overline{\mu}$, while the remaining three rely on the covariance, $\Sigma_{\boldsymbol{\mu}}$, estimated using the standard estimator $\Sigma_{\widehat{\boldsymbol{\mu}}} = \frac{1}{D-1}\sum_{d=1}^{D}\left(\widehat{\boldsymbol{\mu}}_d - \widehat{\overline{\boldsymbol{\mu}}}\right)\left(\widehat{\boldsymbol{\mu}}_d - \widehat{\overline{\boldsymbol{\mu}}}\right)^T$, where $\widehat{\overline{\boldsymbol{\mu}}} = \frac{1}{D}\sum_{d=1}^{D}\widehat{\boldsymbol{\mu}}_d$ is the mean estimate. In order to plot results across K without pure dimension effects displaying prominently, we apply appropriate normalization to each measure that gives them constant scaling in K. We label each measure according to its interpretation as a measure of ideology as follows. The position relative to equal usage (the origin in $\boldsymbol{\mu}$-space) measures how specific the usage is across topics. The size measures how much variability there is in how different individuals discuss that ideology. Eccentricity measures how the correlations limit the expressivity in how the ideology mixes topics. We measure this via reductions in a natural global measure of intrinsic dimension (Recanatesi et al., 2022). Finally, orientedness measures how strongly aligned or anti-aligned pairs of topics are.

Each measure is largely independent of the others and together the set is a largely comprehensive representation of unimodal, non-skewed data, since it covers global statistical features up to second order. We discuss the limitations and missing features of this set of measures in the discussion (Section 4).

3. Results

3.1. Ideology and carbon tax support/opposition

To motivate our study of open-ended responses, we first show how support for the carbon tax depends on the demographic characteristics of the study's respondents. We find that support for carbon tax

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

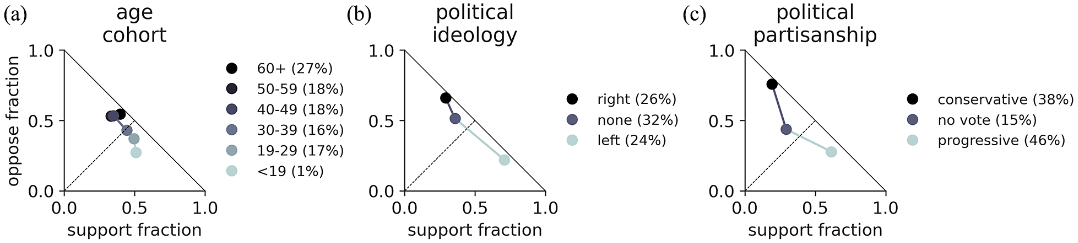

Figure 2. *Carbon tax opposition versus support varies strongly with political demographics. Fraction of respondents that stated they opposed versus the fraction who stated they support the carbon tax over (a) age cohort, (b) political ideology, and (c) political partisanship (measured as voting preference; see Section 2 for details). Addition of the remaining "not sure" fraction (not shown) sums to 1. Opposition is a majority (minority) opinion for dots above (below) the dashed line.*

transitions from a majority to a minority opinion consistently across ideology-related feature dimensions of left-to-right political ideology and liberal-to-conservative partisanship (see Figure 2). These transitions are relatively strong (they extend further away from equal support and opposition) in comparison with a transition with age, shown in Figure 2a as a baseline.

3.2. Discriminability of the word-level representation

Moving onto analysis of the open-ended responses, we first present the response type-conditioned word statistics in Figure 3a. One salient feature is how much the word "tax" dominates in frequency within oppose responses, as compared with the support responses that recruit many frequent words. We then asked what degree word content (which words appear and how many times) could be used to distinguish support and oppose responses. We performed logistic regression on the responses to distinguish support from oppose response type, finding best test accuracy around 80% (Figure 3a; see Section 2 for details about the classification procedure). Word-level features are thus moderately informative in distinguishing support from opposition. This word-level information is spread heterogeneously over the vocabulary. In particular, by rerunning the classification on the n most contributing words (with contribution by

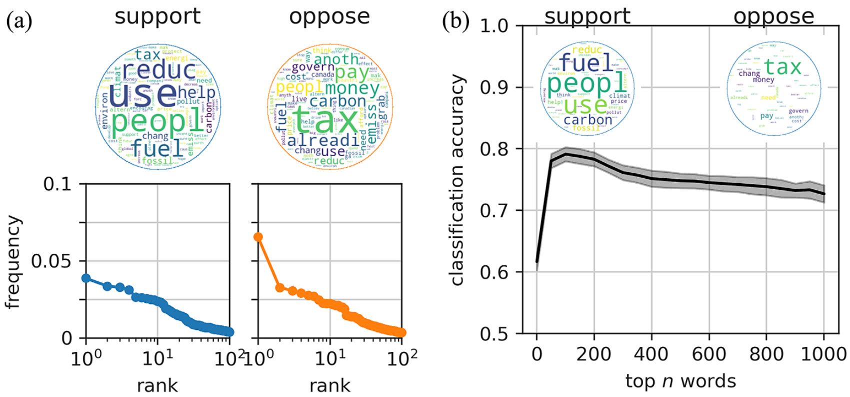

Figure 3. *Word-level representations. (a) Tfidf word cloud (top) and rank plot (bottom) for the word statistics of support and oppose response types. (b) Classification test accuracy of word-level content to predict response type (support/oppose; gray shows ± standard deviation over 100 random train-test splits). Accuracy is plotted for projections of the data onto the n most predictive words (see Section 2 for details). Inset are word clouds of log-transformed frequency-weighted effect size for the $n = 100$ most predictive words for support (left) and oppose (right) classifications.*

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

frequency-weighted effect size) to each of the two target labels (support/oppose), we show in Figure 3b that test accuracy is maximized for about 100 words for each label (200 words overall) out of the more than 2500 stem words appearing in the data. We also show word clouds of the frequency-weighted effect size for $n = 100$ that display the most discriminating words. In sum, many words seem to contribute to a moderate level of discriminability.

A principal drawback of this word-level approach to assessing the discriminating power of the text responses is that individual words are assigned to one or the other response type. Yet, we know the words used by oppose and support responses overlap. For example, “carbon” and “tax” are used in high frequency by both response types. Removing this pair of words allows for higher discriminability (84% accuracy; not shown). That removing these central words increases discriminability suggests first-order word statistics are a poor representation of the semantics of the responses, and that we should pursue approaches that employ latent representations that make use of words in different ways. In particular, it is the context in which these terms are used—what distinct ideas are evoked by the term —that ultimately distinguish the two response types and serves as a better semantic representation.

3.3. Topic modeling

To investigate the contextual origin of response type discriminability using the open-ended responses, we used a generative topic model with latent topic structure. We analyzed two model settings: (1) response type-specific models fit to responses of certain type (support/oppose) and (2) a single model fit to both types of responses combined.

To illustrate the topic content, in Table 2, we show the top words associated with a given topic in a given model for the support and oppose response-type models, as well as for the combined model (all for $K = 7$). A topic’s top words can evoke an unambiguous semantic label for some topics, but not others. For example, topics a and e in the combined model evoke opposition because of the focus on cost of living and because of the word “grab” as in “tax grab,” respectively. However, subjective inspection leaves many cases largely ambiguous about whether a topic in the combined model is weighed more by support or oppose responses. Complementary to top words, top responses can also convey topic quality and even topic labels. For example, we perceived high semantic coherence in topic d, about unfairness of the tax on an international level (as well as, though to a lesser degree, in topic f about economic benefits, in topic c about its effectiveness in bringing about behavior change, etc.). The latter were mostly not transparent in the list of top words, but only in the top responses, that is, in how these words were used. We list some top responses for these topics in the Supplementary Material for reference. It is hard to determine if the ambiguity in determining a topic label from a set of top responses arises from low topic quality reflecting a weakly clustered topic structure in the data, or if it reflects limitations in our ability to identify salient topics. Another convoluting factor is that these topics are specific to the chosen value of K insofar as it imposes a specific number of topics, independent of any natural structure in data. For example, clusters of co-occurring words that are well-separated from each other will nevertheless likely be joined when K is too low and single clusters will likely be labeled incorrectly as multiple distinct clusters when K is too high.

To gain a deeper understanding of the relative changes in the topics as K increased without having to resort to *ad hoc* interpretation of the topics’ top words, we developed two complementary analyses. First, we analyzed how topics breakup as K is increased (we call this a depth phylogeny). Inspired by methods of representing genetic phylogenies, we drew a tree structure obtained from connecting each topic to the topic in the model learned with 1 fewer topic that was nearest in the space of vocabulary word frequencies. In order to see how support and oppose responses spread their use of these topics for topics learned on both responses, we computed a topic depth phylogeny for the combined model (see Figure 4a). Branch thickness denotes closeness.[3] We also added the relative support and oppose contribution to the mixture

[3] We use the inverse Euclidean distance between topics, normalized by the sum of these distances over topics in a given model.

Table 2. *Example topic content (for $K = 7$ and independent topic mixture prior, $\sigma = 1$)*

Topic label	Frequency-ranked words
	Support model
a	chang, environ, climat, pollut, pay, mak, effect, protect, product, earth
b	peopl, help, energi, think, price, better, like, car, decrea, transport
c	emiss, less, increa, caus, govern, anoth, save, solut, drive, option
d	fuel, reduc, fossil, compani, consumpt, must, warm, global, incent, renew
e	tax, money, support, put, good, planet, believ, may, make, work
f	use, need, way, encourag, cost, hope, someth, gener, futur, thing
g	carbon, altern, sourc, find, get, develop, look, forc, creat, corpor
	Oppose model
a	use, fuel, work, car, altern, price, electr, like, stop, better
b	emiss, canada, countri, get, make, world, pollut, less, thing, big
c	carbon, alreadi, live, price, heat, caus, global, high, economi, warm
d	chang, reduc, anoth, grab, climat, noth, effect, believ, hurt, spend
e	tax, much, consum, problem, compani, industri, pocket, year, back, person
f	peopl, money, pay, fuel, need, cost, fossil, way, enough, mak
g	govern, put, think, environ, canadian, drive, emiss, incom, mani, averag
	Combined model
a	pay, alreadi, cost, enough, peopl, know, live, much, work, afford
b	money, mak, way, think, get, use, make, consum, environ, altern
c	fuel, reduc, carbon, chang, use, emiss, fossil, peopl, climat, help
d	canada, carbon, countri, effect, world, problem, pollut, believ, one, heat
e	tax, govern, anoth, grab, noth, put, price, car, canadian, still
f	need, like, support, thing, energi, someth, option, wast, resourc, gener
g	environ, tax, peopl, less, energi, pollut, encourag, altern, use, compani

weight component of that topic as a pie chart at each topic node in the tree. This allowed us to observe the cleaving of more coarse topics at low values of K into their finer-grained internal components as K increased. Reminiscent of phylogenetic analyses, we define a semantic-agnostic topic nomenclature using the row and column labels, respectively. In most cases, we see that topics are recruited by both response types so that single topics alone are insufficient for discriminating oppose from support responses. It is the weighting across multiple topics that gives sufficient information. That said, one salient feature of this particular tree is that topic 7f is the common ancestor for many of the topics at larger values of K. It predominantly loads onto support responses, recapitulating existing results in the literature of a diverse and diffuse set of topics discussed by those who support a carbon tax. In the following section, we will see that this $K = 7$ model is highly discriminative and this topic in particular conferred more than 99% accuracy.

In a second analysis, we assessed the quality of the topics over different K using the standard pair of metrics of exclusivity and semantic coherence (see Section 2). Running the model on each response type separately produced topics whose values on these two quality metrics (Figure 4b) show that topic-averaged values give a linear trade-off between the two, with topic number K setting where along the trade-off the model resides. This suggests the linear combination of semantic coherence and exclusivity (e.g., the variance-minimizing projection) as a scalar metric of topic quality. Across different response types and values of K, the fixed sum (i.e., equal weights) of exclusivity and semantic coherence is highest

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

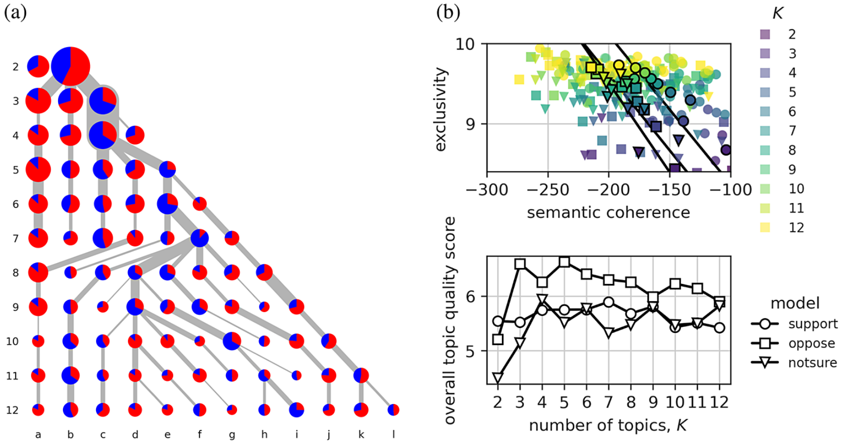

Figure 4. *Topic structure over topic number, K. (a) Topic trees from varying the number of topics, K, in the combined model. A row corresponds to a given value of K starting at the top with $K = 2$ down to $K = 12$. A column refers to one of the topics inferred for that K, indexed by an ordering procedure for visualization. The relative support (blue) and oppose (red) contribution to each response-averaged mixture weight component is shown as a pie chart. Link thickness is inverse Euclidean distance between topics. (b) Topic quality for response type models. Top: Topic exclusivity plotted against topic semantic coherence for topic number $K = 2,\dots,12$ (colors in legend on right) for the three response types (marker type; see legend in Bottom panel). Black outlined markers denote the average over topics obtained for a given K. Bottom: overall score computed from projection orthogonal to average tradeoff line over the three response types (see black lines in top panel). Both plots were made for $\sigma = 0$. Similar results were found for $\sigma = 1$ (see Supplementary Material).*

in *oppose* responses. We attribute this ordering to the singular focus that *oppose* responses have on the word “tax,” as compared to the much more diffuse responses (in word use) of *support* responses. Note, however, that the topic variance in each of the two quality metrics is at least as large as the difference in topic means at fixed topic number, so the aforementioned ranking is only apparent after averaging over topics. Semantic coherence bottomed out at around $K = 10$ topics. Topic quality for the composite model was qualitatively similar.

How predictive of response type is the way topics are recruited in the responses? In Figure 5a, we plot the empirical distribution of mixture weights, projected into its first two principal components, and color-coded with each response type. High separability (especially for $K = 7$) is observed across different topic number K, suggesting high discriminability. We performed a classification analysis on the support/oppose label similar to that using word use and presented at the beginning of our results (see Figure 3b), but now instead using the inferred topic mixture weights. Note that this representation is able to recruit information in the correlations among words that distinguishes the two response types. In Figure 5b, we again show performance as a function of using the most predictive features (here not words, but topic mixture weights). Consistent with the high observed separability, we find accuracies approaching and more than 99% across all values of $K > 2$, demonstrating the power of this latent structure to predict support or opposition to the tax. Accuracies were only marginally lower at 95% or more for $\sigma = 0$ (not shown). For $K = 7$, near 100% accuracy is possible even using only a single topic and perfect accuracy for only two topics. As an example of more fine-grained analysis made possible by our approach, we focus in on these two topics. In Figure 5c, we show the projection of the set of best-estimate mixtures across responses, colored by response type, into the plane in mixture space spanned by the components associated with these two topics. We see that the perfect classification results from topic 7 (topic 5) being strongly suppressed in support (oppose) responses. We further assessed this

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

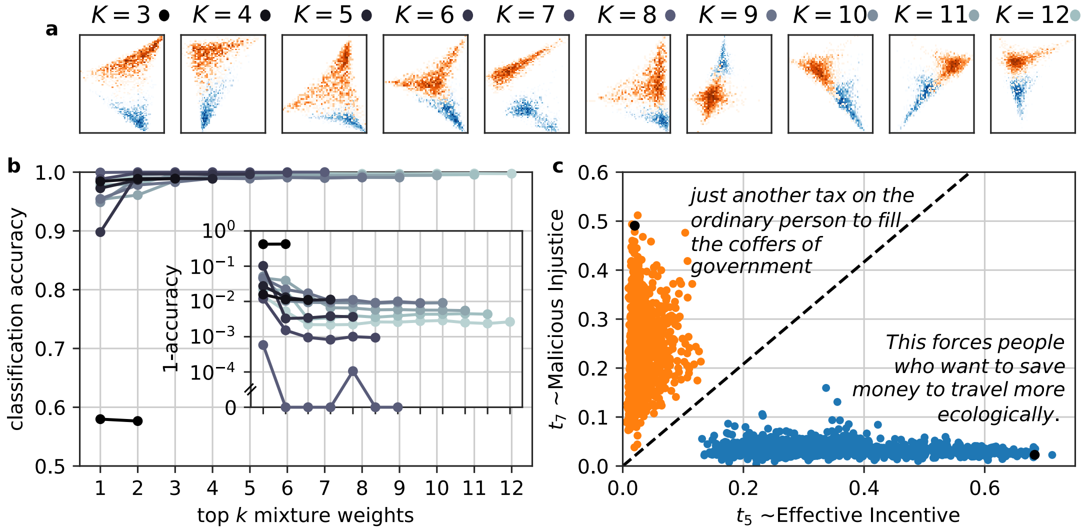

Figure 5. *Topic-level Representations. (a) Projection of inferred mixture weights from the combined model onto the first two PCA components for a range of values of K. A histogram of counts on a fine-grained grid is shown using a white-to-blue and white-to-red color scale for support and oppose responses, respectively. (b) Classification test accuracy for using the best k of the mixture weights ranked by largest frequency-averaged effect size(black to gray is different values of K from 2 to 12). (c) The mixture estimates projected into the two top topics associated with the top k = 2 mixture weights for K = 7. The blacked-dashed line is the optimal classifier that achieves 100% accuracy. The text and mixture position of a top response for each topic is shown as a block dot.*

association by inspecting the top responses for each topic (one from each is shown in Figure 5c). The responses in topic 7 describe the tax as a malicious injustice imposed on citizens, while those in topic 5 describe the effectiveness of the monetary incentive. We note that while the term "tax grab" is the core idea in Topic 7 ("grab" is in 8.5% of oppose responses compared to only 1.5% of support responses), the idea recruits a much wider set of terms operationalized as the support of topic 7 on the vocabulary. It is this wider set of words that confers the discriminability (in this case terms that are not used in high frequency by support responses).

We now turn to the topic-topic correlations to better understand the origin of the topic representation's higher classification performance over word frequency representations. We also arrive at analyses appropriate to investigating the structure of the mixture statistics that putatively hold distinctive features of the underlying ideology. We used the four measures of data cloud geometry summarized in Table 1 to assess the topic-topic correlation structure using $\boldsymbol{\mu}$-representation as explained in the Methods. For the combined response type models, we can directly compare the support and oppose responses at each value of K, so we represent the results as the difference in the values of each measure over the two response types for a range of K around the maximum in the heldout Likelihood (positive values indicate oppose responses exceed support responses on that metric; Figure 6a–d). For the separate response type models, for which the best fitting value of K differs, we simply report both measures over a larger range of K (Figure 6f–i). By showing the heldout likelihood (Figure 6e,j), a comparison at respective maximums could be made. We find consistent results for both comparisons. Namely, oppose responses are more specific, somewhat more variable, much less expressive, and more aligned across a wide range of K around where the held-out likelihood is maximized. All three of the non-weak results are consistent with our hypothesis of a rigid ideology underlying carbon tax opposition.

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

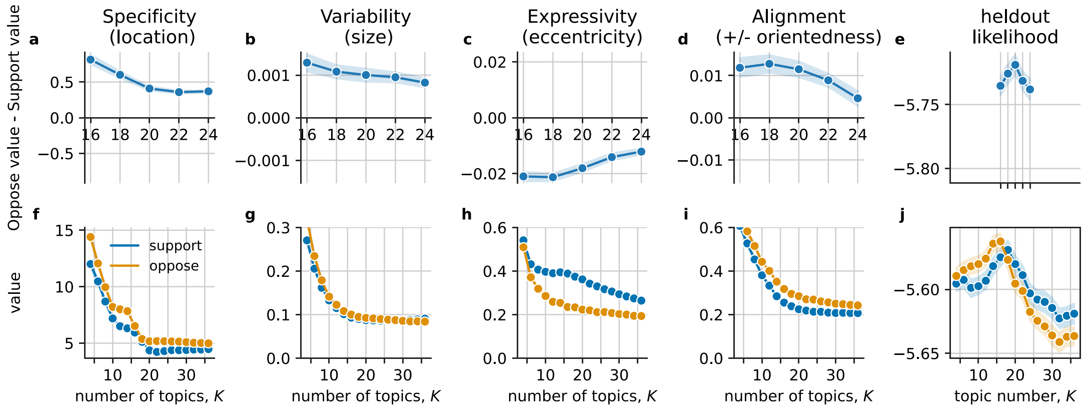

__Figure 6.__ Evaluations of Statistical Measures of Mixture Statistics. Top (a–d): The average difference of the oppose value and the support value in each respective measure from the combined model (equation (4)). Bottom (f–i): The value of the measure for each of the support (blue) and oppose (orange) models (equation (3)). Lines are mean estimates and error bars are the standard deviation of 100 (200) posterior samples for top (bottom). Respective held-out model likelihoods are shown in (e) and (j).

4. Discussion

In this study, we presented a set of principled, quantitative measures and analyses to pin down topic structure in the learned parameters of the STM. We applied them to understand the effects of ideology behind carbon tax opinion in Canada. We find topic mixture weights derived from the learned models are highly predictive of opposition to or support of the carbon tax, and we presented a topic tree analysis similar to phylogenetics to show that this performance arises when the model has a sufficient number of topics that together are discriminating on each response. Finally, we proposed a set of statistical measures on topic mixtures and evaluated them to find that the oppose responses had higher levels of specificity and orientedness, and were less expressive. This suggests carbon tax opposition in Canada is explained by its proponents through a more well-worn, coherent ideology.

How might this result generalize to carbon tax opposition in other countries? There is some debate in the literature regarding the generalizability of ideology structuring climate change beliefs (Lewis et al., 2019). When looking at environmental taxes specifically, ideology plays an important role, but is mediated by the quality of government (or at least public perception of it), such that progressives in low quality of government contexts are not more likely to support environmental taxation (Davidovic et al., 2020). Declining public faith in government may then hinder public acceptance of fiscal policy on climate change, independent of ideology.

We anticipate the methods we present here can be useful to social scientists who are interested in inferring social and political phenomenon from open-ended survey responses. Most obviously, our methods could be used to further quantify the structure of collective beliefs. In a longitudinal approach, they could also reveal the dynamics of these collective beliefs, and in particular their response characteristics, for example from large-scale interventions or events (e.g., the COVID-19 pandemic, Russia-Ukraine war). A study of the timescales over which the effects of interventions decay would inform the recent work on the notion of inoculation to misinformation. To push the metaphor, the characteristics of the memory it induces have not yet been deeply probed (Roozenbeek et al., 2022). Our work suggests that the topic neighborhood around the focus of any given intervention may play a role via the topic-topic correlation structure. This could also impact policy design, for example in how the carbon tax is promoted. Typical communications advice is to focus on a single issue/message to not overload the target audience. However, neighborhood effects that are strongly restoring (Hall and Bialek, 2019) could rapidly erase the effects of single-issue interventions. Instead, a communications strategy aimed at a local neighborhood of

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

target issues in some appropriately defined issues space may be more effective at loosening the grip that ideology has on opinion.

There are many potential directions in which to take this work. For one, the existing set of metrics of mixture statistics only cover first and second-order statistical moments. Additional low-order statistics could be included, for example, how concentrated is the distribution around its center (the expectation of $\|\boldsymbol{\mu}\|$ might provide signal into how hollow is the data cloud). In another direction, the efficiency of the topic representations suggests it could serve as an input embedding space for use in deep learning approaches to large-scale social computations (e.g., those run on large social media platforms). For example, recent work on reinforcement learning with human feedback for large language models has focused on generating consensus statements from a set of human-generated opinions (Bakker et al., 2022). A currently limiting constraint of this approach is how many human-generated text inputs can be used. A topic space learned from many opinions could circumscribe this constraint.

We probed model behavior over a range of K around the maximum in held-out model likelihood. We were able to make conclusions from the fact that the sign of the difference of the values of the metrics was unchanged over these ranges. This will not be true in general, in which case one approach is to marginalize K out of the problem. In a fully Bayesian approach, given a prior over K, one would infer a posterior over K and take the posterior average over any K-dependent quantities, for example, average topic quality or the statistical metrics. Given that models with different K have a different number of parameters, a complexity-penalty term such as the AIC could be added to the posterior to account for variable model complexity.

One direction we did not pursue here was exploring how different meta-data models affect the results. Are there specific factor models of fewer covariates that best explain the data? As we motivated our choice of the all-in model, this is a delicate analysis requiring disentangling various cases. For example, is car use a confounder or collider of the effect of residence environment? Such distinctions will influence estimated effect sizes (Wysocki et al., 2022). A first step along this direction is to further analyze the results in the case of a sparsity prior on Γ. Including meta-data model dependence in topic content is a related direction for future work. Note that in the absence of rich metadata, the STM model reduces roughly to the Correlated Topic Model (Blei and Lafferty, 2005), in which case topic correlations can be studied via the mixture covariance matrix parameter, $\Sigma_{\boldsymbol{\theta}}$.

Finally, there is a broader question of the validity of the generative model class. In particular, are topic models with latent structure suitable mathematical representations of the semantic content of networks of beliefs? How much is lost when representing language as text, without syntax and without positional encoding? How much information about beliefs is contained in language, given that a word can have one of many different meanings depending on who is speaking (Marti et al., 2023). These questions broach broader philosophical issues around meaning in natural human language. The rich semantics in (position-encoding) large language models that achieve in-context learning (Mollo and Millière, 2023), suggest text is a sufficiently rich representation of natural language from which to extract semantics. In fact, to the extent that semantics are not position-encoded (so-called exchangeability), the autoregressive objective on which these models are trained is formally equivalent to the posterior inference on a topic model (Zhang et al., 2023). Stripping text of its positional encoding must result in some loss of meaning, though we note that it need not degrade and may even improve performance in some downstream tasks, for example, Kazemnejad et al. (2023). We see topic models as complementing more powerful large language models by having more interpretable latent spaces, and we exploit that interpretability in this work. We hope it inspires the development, refinement, and application of topic models in computational data science and beyond.

Acknowledgments. The authors would like to acknowledge helpful discussions with Matto Mildenberger, Kathryn Harrison, and Dhanya Sridhar.

Author contribution. M.P.T.: Conceptualization (Equal), Formal analysis (Lead), Investigation (Lead), Methodology (Lead), Software (Lead), Writing—original draft (Lead), Writing—review and editing (Equal). E.L.: Conceptualization (Equal), Data curation (Lead), Supervision (Lead), Investigation (Supporting), Writing—review and editing (Equal).

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

Competing interest. The authors declare none.

Data availability statement. The raw survey data obtained for Mildenberger et al. (2022) was not made publicaly available for privacy reasons (they include some derivative data on the Harvard Dataverse replication archive at https://doi.org/10.7910/DVN/3WBCH9). For this publication, we created a privacy-preserving processed dataset of only the fields we used in our analysis. We have stored this dataset in a file that we make publically available in a GitHub repository (https://github.com/mptouzel/PuelmaTouzel_EDS_2024_carbontax), alongside the Python scripts and computational notebooks needed to reproduce all figures of the paper.

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. M.P.T. acknowledges the support from the Canada CIFAR AI Chair Program and the Canada Excellence Research Chairs Program via Dr. Irina Rish. Funding was also drawn from a Social Sciences and Humanities Research Council of Canada award (#435-2017-1388) in which E.L. is the principal investigator.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/eds.2023.44.

References

Aitchison J (1982) The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological) 440*(2), 139–177. Available at http://www.jstor.org/stable/2345821.

Aldy JE and Stavins RN (2012) The promise and problems of pricing carbon: Theory and experience. *The Journal of Environment & Development 210*(2), 152–180.

Bakker MA, **Chadwick MJ**, **Sheahan HR**, **Tessler MH**, **Campbell-Gillingham L**, **Balaguer J**, **McAleese N**, **Glaese A**, **Aslanides J**, **Botvinick MM and Summerfield C** (2022) Fine-tuning language models to find agreement among humans with diverse preferences. Available at https://arxiv.org/abs/2211.15006.

Baldassarri D and Gelman A (2008) Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology 1140*(2), 408–446.

Baranzini A, **Goldemberg J and Speck S** (2000) A future for carbon taxes. *Ecological Economics 320*(3), 395–412. https://doi.org/10.1016/S0921-8009(99)00122-6. Available at https://www.sciencedirect.com/science/article/pii/S0921800999001226.

Birch S (2020) Political polarization and environmental attitudes: A cross-national analysis. *Environmental Politics 290*(4), 697–718.

Blei DM and Lafferty JD (2005) Correlated topic models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems, NIPS'05*. Cambridge, MA: MIT Press, pp. 147–154.

Dalton RJ and Wattenberg MP (2002) *Parties without Partisans: Political Change in Advanced Industrial Democracies*. Oxford: Oxford University Press. https://doi.org/10.1093/0199253099.001.0001.

Davidovic D, **Harring N and Jagers SC** (2020) The contingent effects of environmental concern and ideology: Institutional context and people's willingness to pay environmental taxes. *Environmental Politics 290*(4), 674–696. https://doi.org/10.1080/09644016.2019.1606882.

Drews S and Van den Bergh JC (2016) What explains public support for climate policies? A review of empirical and experimental studies. *Climate Policy 160*(7), 855–876.

Eisenstein J, **Ahmed A and Xing EP** (2011) Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. Madison, WI: Omnipress, pp. 1041–1048.

Gershman SJ (2019) How to never be wrong. *Psychonomic Bulletin and Review 260*(1), 13–28. https://doi.org/10.3758/s13423-018-1488-8.

Grimmer J, **Roberts M and Stewart B** (2022) *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press. Available at https://books.google.ca/books?id=cL40EAAAQBAJ.

Haites E (2018) Carbon taxes and greenhouse gas emissions trading systems: What have we learned? *Climate Policy 180*(8), 955–966.

Hall G and Bialek W (2019) The statistical mechanics of twitter communities. *Journal of Statistical Mechanics: Theory and Experiment 20190*(9), 093406. https://doi.org/10.1088/1742-5468/ab3af0.

Kalmoe NP (2020) Uses and abuses of ideology in political psychology. *Political Psychology 410*(4), 771–793.

Kazemnejad A, **Padhi I**, **Ramamurthy KN**, **Das P and Reddy S** (2023) The impact of positional encoding on length generalization in transformers.

Klenert D, **Mattauch L**, **Combet E**, **Edenhofer O**, **Hepburn C**, **Rafaty R and Stern N** (2018) Making carbon pricing work for citizens. *Nature Climate Change 80*(8), 669–677. https://doi.org/10.1038/s41558-018-0201-2.

Lachapelle E (2017) Communicating about carbon taxes and emissions trading programs. In *Oxford Research Encyclopedia of Climate Science*. Oxford: Oxford University Press.

Lakoff G (2010) Why it matters how we frame the environment. *Environmental Communication 40*(1), 70–81. https://doi.org/10.1080/17524030903529749.

https://doi.org/10.1017/eds.2023.44 Published online by Cambridge University Press

Lewis GB, **Palm R and Feng B** (2019) Cross-national variation in determinants of climate change concern. *Environmental Politics 280*(5), 793–821. https://doi.org/10.1080/09644016.2018.1512261.

Marti L, **Wu S**, **Piantadosi ST and Kidd C** (2023) Latent diversity in human concepts. *Open Mind 7*, 79–92. https://doi.org/10.1162/opmi_a_00072.

McCright AM and Dunlap RE (2011) The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *The Sociological Quarterly 520*(2), 155–194.

McCright AM, **Dunlap RE and Marquart-Pyatt ST** (2016) Political ideology and views about climate change in the European union. *Environmental Politics 250*(2), 338–358.

McGlade C and Ekins P (2015) The geographical distribution of fossil fuels unused when limiting global warming to 2 c. *Nature 5170*(7533), 187–190.

Mildenberger M, **Lachapelle E**, **Harrison K and Stadelmann-Steffen I** (2022) Limited impacts of carbon tax rebate programmes on public support for carbon pricing. *Nature Climate Change 120*(2), 141–147. https://doi.org/10.1038/s41558-021-01268-3.

Mollo DC and Millière R (2023) The vector grounding problem.

Pawlowsky-Glahn V and Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment 150*(5), 384–398. https://doi.org/10.1007/s004770100077.

Povitkina M, **Jagers SC**, **Matti S and Martinsson J** (2021) Why are carbon taxes unfair? Disentangling public perceptions of fairness. *Global Environmental Change 700*, 102356. https://doi.org/10.1016/j.gloenvcha.2021.102356.

Rabe BG (2018) *Can we Price Carbon?* Cambridge, MA: MIT Press.

Recanatesi S, **Bradde S**, **Balasubramanian V**, **Steinmetz NA, and Shea-Brown E** (2022) A scale-dependent measure of system dimensionality. *Patterns 30*(8), 2666–3899. https://doi.org/10.1016/j.patter.2022.100555.

Roberts ME, **Stewart BM and Tingley D** (2019) Stm: An R package for structural topic models. *Journal of Statistical Software 91*, 1–40.

Roberts ME, **Stewart BM**, **Tingley D**, **Lucas C**, **Leder J-Luis**, **Gadarian SK**, **Albertson B and Rand DG** (2014) Structural topic models for open-ended survey responses. *American Journal of Political Science 580*(4), 1064–1082. https://doi.org/10.1111/ajps.12103. Available at http://dvn.iq.harvard.edu/dvn/dv/ajps.

Roozenbeek J, **van der Linden S**, **Goldberg B**, **Rathje S and Lewandowsky S** (2022) Psychological inoculation improves resilience against misinformation on social media. *Science Advances 80*(34), eabo6254. https://doi.org/10.1126/sciadv.abo6254. Available at https://www.science.org/doi/abs/10.1126/sciadv.abo6254.

Savin I, **Drews S**, **Maestre-Andrés S and van den Bergh J** (2020) Public views on carbon taxation and its fairness: A computational-linguistics analysis. *Climatic Change 1620*(4), 2107–2138. https://doi.org/10.1007/s10584-020-02842-y.

Tvinnereim E, **Fløttum K**, **Gjerstad Ø**, **Johannesson MP and Nordø ÅD** (2017) Citizens' preferences for tackling climate change. Quantitative and qualitative analyses of their freely formulated solutions. *Global Environmental Change 460*, 34–41. https://doi.org/10.1016/j.gloenvcha.2017.06.005.

Welsby D, **Price J**, **Pye S and Ekins P** (2021) Unextractable fossil fuels in a 1.5Â° c world. *Nature 5970*(7875), 230–234.

Wysocki AC, **Lawson KM and Rhemtulla M** (2022) Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science 50*(2),25152459221095823. https://doi.org/10.1177/25152459221095823.

Zhang L, **McCoy RT**, **Sumers TR**, **Zhu J-Q**, **Griffiths TL** (2023) Deep de Finetti: Recovering topic distributions from large language models. Available at arXiv:2312.14226.

Cite this article: Puelma Touzel M and Lachapelle E (2024). Ideology from topic mixture statistics: inference method and example application to carbon tax public opinion. *Environmental Data Science*, 3: e10. doi:10.1017/eds.2023.44