Supporting Information: *Measuring Agenda Setting in Interactive Political Communications*

Contents

A SAMPLER

In all simulations and applications, I estimate SITS using Markov chain Monte Carlo, specifically, a Gibbs sampler written in Java by Viet An Nguyen (Nguyen 2014). The collapsed Gibbs sampler has similarities to the collapsed Gibbs sampler for LDA (Griffiths and Steyvers 2004). The latent topic distributions ($\theta_{d,t}$) and topics (ϕ_k) have been integrated out of the full conditional probabilities for $z_{d,t,n}$ and $l_{d,t}$, and these parameters are estimated using the posterior distributions of topic assignments. Similarly, speaker agenda setting measures (π_m) are integrated out of the full conditional probabilities and are estimated from the posterior distributions of topic changing indicators ($l_{d,t}$). Thus, an iteration of the sampler samples the topics assigned to each word in a speaking turn ($z_{d,t,n}$) as well as the topic shift indicator assigned to each turn ($l_{d,t}$). Full details for the sampling equations can be found in the Appendix of Nguyen et al. (2014).

The starting values for $z_{d,t,n}$ are randomly assigned a value $[1, K]$. The starting values for $l_{d,t}$ are randomly assigned a starting value $[0, 1]$. For each application, I ran three chains, each for 500,000, discarding the first 400,000 as burn-in iterations. Because of the high-dimensional nature of the topic assignment parameters ($z_{d,t,n}$), I retain only every 100^{th} iteration, and I then use the modal sampled value as my estimate of each $z_{d,t,n}$. I did not estimate $l_{d,t}$ parameters when the speaking turn has less than 5 tokens unless otherwise noted.

B TEXT PREPROCESSING

Unfortunately, existing software, procedures, and metrics for assessing topics models is not always easily applied to SITS and a corpus of interactions. This is because part of the inferential goal of SITS is to infer coherent segments within a corpus of interactions. In other words, what to consider a “document” or the relevant “unit of analysis“ in the case of interactions is not clear. Within a corpus of interactions, the text is organized two ways. First, the text is organized at the interaction-level. So to use existing metrics and software, a researcher could define a“document” at the interaction-level. However this would ignore that the set of topics receiving attention shifts as the interaction unfolds by obscuring all speaking turns into a single instance of text. The second way a researcher could make use of existing metrics and software is by utilizing that the text in an interaction is also organized by speaking-turns. Therefore, a “document” could be defined at the speaking-tun level, however, the resulting texts would often be very short and it would be difficult to assess differences (e.g., Denny and Spirling 2018) or infer coherent topics (e.g., Nikita 2020) with such short texts.

Nevertheless, I use the `preText` R package and methodology introduced by Denny and Spirling (2018) to assess the potential implications of the preprocessing decisions they highlight. Needing to choose how to represent the text in order to use their method—at the speaking-turn or interaction level—I chose to represent the text and the interaction-level as the speaking turns in my corpora were too short to calculate meaningful differences. However, I am not actually modeling the text at the interaction-level. Therefore, I take the results from the `preText` analysis as suggestive of ways in which my results may be sensitive to preprocessing decisions. I replicate my substantive results with the alternative preprocessing steps when suggested by `preText` results. The specific preprocessing steps taken for each corpus are outlined in Appendix B.1 (presidential debates), Appendix B.2 (in-person deliberations), and Appendix B.3 (online discussions).

B.1 Presidential debates

To consider the implications of preprocessing choices for the presidential debates corpus, I used `preText` as discussed above. Because the coefficient on removing stopwords is significantly different from zero, and I have no theoretical justification for a preferred choice on this preprocessing step, I replicate the analysis with and without stopwords as Denny and Spirling (2018) advise. The coefficient on removing punctuation is also significant and has a larger coefficient. This is likely due to the fact that punctuation was inconsistently included in the transcripts I used across the 7 election cycles. Therefore, I chose to remove punctuation and I do not replicate results when retaining it.

Specifically, I preprocess the text by removing punctuation, removing numbers, transforming all words to lower case, stemming, and removing a set of stopwords that included common English stopwords, annotations to the transcripts (such as "applause"), and the name and titles of all speakers. Finally, I removed infrequent terms that occurred in fewer than 0.5% of speaking turns. I have two theoretical motivations in mind when removing these terms the debate text. First, because the debates span 24 years, the debates may feature different specific policies across time (e.g., No Child Left Behind) despite discussing similar issues regardless of the year (e.g., education). Therefore, I removed infrequent terms from the corpus with the intention to remove several terms that are election-specific. Second, the moderators and candidates often refer to each others' names and titles. I remove these terms because I certain speakers' names may co-occur with other terms in ways that do not help me explore the general topics across the set of presidential debates.

There were 20 debates across the 7 elections. On average, there were 190 speaking turns per debate. After preprocessing steps including the removal of stopwords, the corpus vocabulary then contained 1015 terms across 3,818 speaking turns. When removing stopwords, the corpus vocabulary then contained 944 unique terms. Appendix C details model selection for the two preprocessed corpora. Appendix E details the replication results.

B.2 In-person deliberations

There were 10 deliberations regarding the BYU Dress and Grooming Standards. On average, there were 90 speaking turns per deliberation. Because the deliberations were not professionally transcribed, capitalization, punctuation, and numbers were used inconsistently. Therefore, to preprocess the text, I started by removing all capitalization, punctuation, and numbers. I used `preText`, as discussed above, to consider the implications of stemming, removing stopwords, and removing infrequent terms. Because the coefficient on removing stopwords is significantly different from zero and I have no theoretical justification for a preferred choice on this preprocessing step, I replicate the analysis with and without stopwords as Denny and Spirling (2018) advise.

After preprocessing steps including the removal of stopwords, the corpus vocabulary then contained 776 words across 899 speaking turns. When removing stopwords, the corpus vocabulary then contained 667 unique terms. Appendix C details model selection for the two preprocessed corpora. Appendix F details the replication results.

B.3 Online discussions

There were 91 discussions. On average, there were 16 speaking turns per discussion. To preprocess the text, I began by transforming all words to lowercase as capitalization was used inconsistently by the participants. I also selectively removed punctuation, keeping exclamation marks, question marks, and punctuation meant to mimic emojis (e.g., ":)"). I kept this punctuation because it was an important part of the communication in an online environment. Further, I cleaned the charity names so all variations of how participants referred to a charity are referred to by the same token. For this application, I kept numbers because participants often appealed to the charities' scores on Charity Navigator (see Appendix G.2 for more information). I used preText, as discussed above, to consider if results may be sensitive to preprocessing steps. Results suggest that removing punctuation results in significant differences in the texts. However, I am motivated to selectively keep punctuation because I think it is an important part in distinguishing different topics within the online environment. For example, when greeting one's partner, a participant may use exclamation marks. However, when discussing the charities, they may use percentage signs when talking about their scores on Charity Navigator.

After these preprocessing steps, the corpus vocabulary then contained 523 terms across 1516 speaking turns.

C HYPERPARAMETERS AND MODEL SELECTION

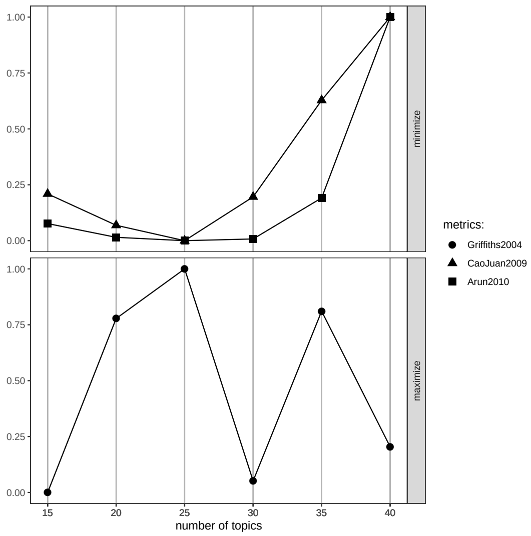
I chose hyperparameters for modeling each corpus based on popular recommendations (and often, default values in statistical softwares) in the literature for parameterizing LDA. Broadly, the advice is to induce sparsity in the the document-topic and topic-word distributions with choice of $\alpha < 1$ and $\beta < 1$, respectively. The intuition is that smaller α represents a prior belief that only a few topics make up large proportions of each document, rather than all topics being represented equally. Likewise, smaller β represents a prior belief that only a few words have high probability for each topic, rather than many words being strong indicators of the topic. In practice, the advice is to let $\alpha = 1/K$ and $\beta = .1$ Griffiths and Steyvers (2004), and I follow this guidance for all applications of the SITS model. Because there is not advice to appeal to in the literature, I choose a noninformative, uniform prior over speaker agenda setting parameters by letting $\gamma = 1$ in all cases.

Finally, for each application, I must choose the number of topics K. I follow the strategy of Griffiths and Steyvers (2004), which is to fix the values of α and β as just discussed. Then, I can assess the consequences of varying the number of topics, K. As Griffiths and Steyvers note, in this sense, choosing K becomes a question of model selection. I use the ldatuning R software which calculates metrics proposed in the literature to choose a preferable value of K for the corpus (Nikita 2020). However, as discussed in regard to preprocessing the text in Appendix B, existing software and metrics for assessing topics models is not easily applied to SITS. Therefore, to choose K, I aggregate the text from all speaking turns within an interaction into one document. Then, with fixed $\alpha = 1/K$ and $\beta = .1$, I use the available metrics to choose a preferable K for the corpus. The specific choice of K is described in Appendix C.1 (validation exercises), Appendix C.2 (presidential debates), Appendix C.3 (in-person deliberations), and Appendix C.4 (online discussions).

C.1 Validation exercises

Figure 1 visualizes model selection results for the validation exercise using the 1992, 2004, and 2008 presidential debates. The available metrics suggest choosing $K = 25$.

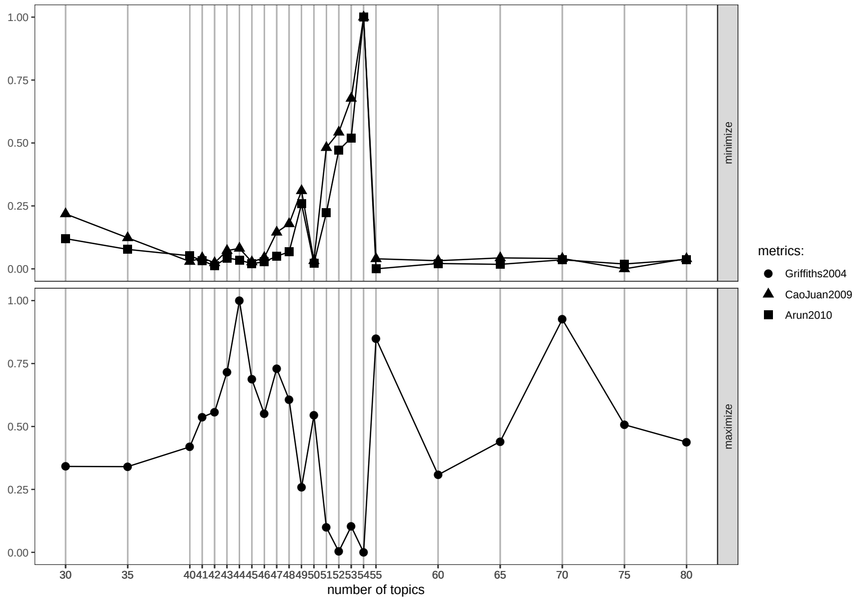
Figure 1: Choosing a preferable K for topics and segments validation exercise



C.2 Presidential debates

Figure 2 visualizes model selection results. When removing stopwords, the available metrics suggest choosing $K = 44$. When keeping stopwords, the metrics suggest $K = 48$.

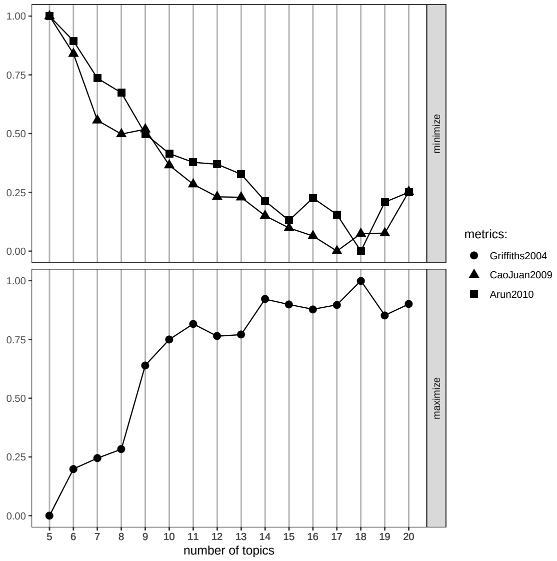
Figure 2: Choosing a preferable K for presidential debates corpus



C.3 In-person deliberations

Figure 3 visualizes model selection results. When removing stopwords, the available metrics suggest choosing $K = 18$. When keeping stopwords, the metrics suggest $K = 20$.

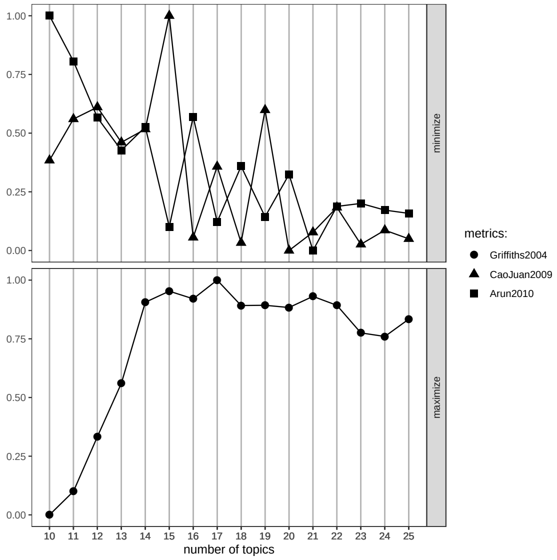
Figure 3: Choosing a preferable K for in-person deliberation corpus



C.4 Online discussions

Figure 4 visualizes model selection results, and the available metrics suggest choosing $K = 17$.

Figure 4: Choosing a preferable K for online discussions corpus



D VALIDATION EXERCISES

D.1 Latent topic shifts

D.1.1 Study design

The researcher provided topics were the following: their favorite class; their partisanship; if they were an in-state or out-of-state student; the fairness of tuition at their institution; whether illegal immigrants attending high school in the US should qualify for federal financial aid for college; if there is too much emphasis on standardized testing in high school; and if their institution should drop standardized testing scores as an application criteria.

D.1.2 Comparison to other methods

As an added validation exercise, I assess if other automated text analysis methods can be used to infer where topic shifts occur within texts arising from interactions. Automated text analysis has not yet been used in political science to identify locations *within* a text where latent shifts in topic may occur. Automated text analysis has been used to judge similarity or difference *across* texts, but in these instances, the relevant boundaries of texts are already defined. One purpose of SITS is to find these relevant boundaries within an interaction by estimating where topic shifts occur. To validate this, I utilize a corpus of interactions with exogenous topic shifts. I show that standard text analysis methods used in political science to assess text difference do not perform well when adapted to the task of finding topic shifts within an interaction.

Recall, SITS estimates a binary topic shift variable for each speaking turn. I averaged the posterior mean for each turn-level variable for each model, thus calculating the probability that each turn shifts topic. To assess if SITS can accurately identify where latent shifts in topic do (and do not) occur, I plot receiver operating characteristic (ROC) curves in Figure 5. ROC curves are a visualization of the diagnostic ability of a binary classifier—in this case, classifying turns as topics shifts or not—while varying the threshold at which to determine classification—in this case, varying the probability (i.e., 0.5 or 0.75) at which a turn is considered a topic shift.

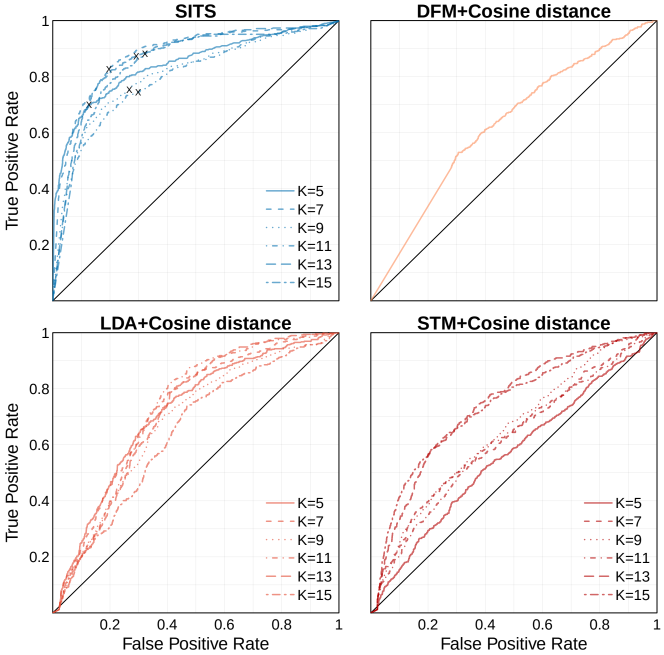
The upper left plot in Figure 5 shows the classification ability across a range of K values to assess if the performance is robust to choice of K. SITS has a high true positive rate (x-axis) and a lower false positive rate (y-axis) for each value of K. Crosses indicate the diagnostic ability of each model when classifying turns posterior probability of $P(l_{d,t}) \geq .5$ as shifts. At this threshold, SITS classifies about 80% of true topic changes correctly when $K = 7$, for example.

Because one might think simpler automated text analysis methods could be leveraged to measure agenda setting, I compare the performance of SITS to standard text analysis methods used in political science. Arguably, a simpler way to identify topic shifts could be to measure dissimilarity of consecutive speaking turns. Then if the text in turn t is dissimilar to the text of turn $t-1$, it is likely the speaker of turn t shifted the topic.

Detecting similarity between two texts is a difficult task due to the high-dimensional nature of text data, making this an active research area in political science (e.g., Mozer et al. Forthcoming). To attempt to detect turns that change the topic, I represent the text of a speaking turn in three ways: by its bag-of-words term frequencies and by its latent topic proportions estimated by an LDA and a Structual Topic Models (STM) topic model (Roberts et al. 2014). I then assess the classification of topic shifts for each of these representations of the text with cosine distance, a commonly used distance metric for text in political science.

Since the cosine distance metric does not have an intuitive threshold for determining if a topic shift occurred or not, I again use ROC plots to consider classification at any given value of the

Figure 5: Topic-shift classification accuracy of SITS vs standard automated text methods



Note: Figure presents ROC curves for turn-level topic shift classification. Standard text as data methods do hardly better or worse than random guessing (indicated by diagonal gray line), regardless of chosen threshold when adapted to the task of identifying topic changes within an interactive setting.

distance metric. Figure 5 plots the ROC curves. The second pane shows the classification ability of representing the texts with frequencies from a document-feature matrix (DFM) and using cosine distance to assess the difference between consecutive speaking turns. This approach barely outperforms random guessing, as indicated by the diagonal black line.

The third pane of Figure 5 shows classification results when representing the text with topic proportions from LDA topic models. I again plot a subset of the models estimated with varying values of K. This approach—assessing the difference between estimated topic proportions—performs better than using frequencies from the document-feature matrix. However, this approach fails to accurately discriminate true positives from false positives. When this approach identifies 80% of true topic shifts correctly, it also incorrectly identifies 50% of non-shifts as shifts.

The final pane of Figure 5 shows classification results when representing the text with using topic proportions from STM topic models estimated with a prevalence covariate to indicate which discussion the text came from. This approach actually performs worse than LDA, barely outperforming random guessing.

Taken together, this validation exercise suggests that identifying speaking turns that change the topic cannot be reliably accomplished with commonly used automated text analysis methods; however, SITS can accurately identify the speaking turns that should be attributed with shifts in the agenda.

D.2 Latent topics

D.2.1 MTurk task

Amazon Mechanical Turk (MTurk) Workers were paid $0.10 per completed task. Participants were paid the agreed compensation within the MTurk platform within hours of completing the task. To qualify for the work, they had to be 18 years or older, located in the United States, hold a 99 percent or higher task approval rating, and have completed 2000 or more tasks. In addition, they had to take a qualification test and correctly answer two simple, contrived topic intrusion tasks to ensure they understood the task at hand. There were 38 unique Workers who completed ratings. Workers on average completed 10 tasks. The instructions for the Amazon Mechanical Turk Workers were to "please indicate the set of words that is most unrelated to the passage."

D.3 Latent segments

D.3.1 Additional details for inferring topic shifts and clustering segments

Since the hand-coded data was at the phrase-level, I aggregated these codings at the turn-level by calculated a continuous measure of the fraction of phrases that were off topic within a turn. I estimated SITS after preprocessing the text by removing punctuation, removing numbers, transforming all words to lower case, stemming, and removing a set of stopwords that included common English stopwords, annotations to the transcripts (such as "applause"), and the name and titles of all speakers. Finally, I removed the 35 terms that occurred in more than 10% of speaking turns and I removed infrequent terms that occurred in fewer than 0.5% of speaking turns. The resulting document-feature matrix had 4,004 speaking turns and 1,074 features.[1]

For each method, I binarized the continuous topic-shift measure. For SITS, I considered a speaking turn to change the topic if the posterior mean of the latent topic shift indicator was greater than or equal to .5. Likewise, for the hand-coded measure, I considered a speaking turn to change the topic if over half of the turn's phrases were deemed off topic.[2]

The second step was to determine the similarity of these segments (i.e., to cluster the segments). For the hand-coded segments, I considered two segments to be similar if the most frequently assigned topic within the segment was the same. To take a comparable approach with SITS, I considered two segments to be similar if the largest component of the two segments' estimated topic proportions was the same.

D.3.2 MTurk task

Amazon Mechanical Turk (MTurk) Workers were paid $0.20 per completed task (rating one pair of segments). Participants were paid the agreed compensation within the MTurk platform within hours of completing the task. To qualify for the work, they had to be 18 years or older, located in the United States, hold a 99 percent or higher task approval rating, and have completed 2000 or more tasks. In addition, they had to take a qualification test and correctly answer two simple, contrived pairings to ensure they understood the task at hand. The instructions for the Amazon Mechanical Turk Workers were the following:

> Please rate the similarity of ***the general topics*** being discussed within the two texts, such as education, immigration, elections, or defense.
>
> These texts are segments from the 1992, 2004, or 2008 United States general election presidential debates. These debates feature the main election candidates, a moderator, and sometimes audience or panelist members posing questions.

[1] Appendix C details my approach to choosing hyperparameter values and model selection.

[2] For the hand-coded data, I also considered turns that had 20 or fewer words as unable to change the topics, as these turns were often coded as changing the topic even though the candidate was not changing the substantive topic of interest but instead was trying to interject, for example.

> Since these texts come from many different United States presidential debates, ***please do not base your ratings on if specific policies, candidates, or time-periods differ***.

There were 31 unique Workers who completed ratings. Workers on average rated 12 pairs and were limited to rating no more than 50 of the 100 pairs.

D.3.3 Example pairing

The following is an example of a pair of SITS segments drawn from different topics. For this example, all Workers unanimously said the general topics being discussed in these two pieces of text were *unrelated*. (For reference, the main topic for Document 1 was "financial crises" and had the following FREX top words: financi, crisi, bank, street, packag, home, treasuri, problem. The main topic for Document 2 was "families" and had the following FREX top words: love, children, famili, dream, mother, person, women, daughter.)

Document 1

MCCAIN: No, I – look, we've got to fix the system We've got fundamental problems in the system And Main Street is paying a penalty for the excesses and greed in Washington, DC, and on Wall Street So there's no doubt that we have a long way to go And, obviously, stricter interpretation and consolidation of the various regulatory agencies that weren't doing their job, that has brought on this crisis But I have a fundamental belief in the goodness and strength of the American worker And the American worker is the most productive, the most innovative America is still the greatest producer, exporter and importer But we've got to get through these times but I have a fundamental belief in the United States of America And I still believe, under the right leadership, our best days are ahead of us

LEHRER: All right, let's go to the next lead question, which is essentially following up on this same subject And you get two minutes to begin with, Senator McCain are there fundamental differences between your approach and Senator Obama's approach to what you would do as president to lead this country out of the financial crisis?

Document 2

CLINTON: A family involves at least one parent, whether natural or adoptive or foster, and children. A good family is a place where love and discipline and good values are transmuted (sic) from the elders to the children, a place where people turn for refuge, and where they know they're the most important people in the world. America has a lot of families that are in trouble today. There's been a lot of talk about family values in this campaign. I know a lot about that. I was born to a widowed mother who gave me family values, and grandparents. I've seen the family values of my people in Arkansas. I've seen the family values of all these people in America who are out there killing themselves working harder for less in a country that's had the worst economic years in 50 years and the first decline in industrial production ever. I think the president owes it to family values to show that he values America's families, whether they're people on welfare you're trying to move from welfare to work, the working poor whom I think deserve a tax break to lift them above poverty if they've got a child in the house and working 40 hours a week, working families who deserve a fair tax system and the opportunity for constant retraining; they deserve a strong economy. And I think they deserve a family and medical leave act. Seventy two other nations have been able to do it. Mr. Bush vetoed it twice because he says we can't do something sevenety two other countries do, even though there was a small business exemption. So with all the talk about family values, I know about family values – I wouldn't be here without them. The best expression of my family values is that tonight's my 17th wedding anniversary, and I'd like to close my question by just wishing my wife a happy anniversary, and thank you, my daughter, for being there.

LEHRER: President Bush, one minute.

BUSH: Well, I would say that one meeting that made a profound impression on me was when the mayors of the big cities, including the mayor of Los Angeles, a Democrat, came to see me, and they unanimously said the decline in urban America stems f So I do think we need to strengthen family. When Barbara holds an AIDS baby, she's showing a certain compassion for family; when she reads to children, the same thing. I believe that discipline and respect for the law – all of these things should be taught to children, not in our schools, but families have to do that. I'm appalled at the highest outrageous numbers of divorces – it happens in families, it's happened in ours. But it's gotten too much. And I just think that we ought to do everything we can to respect the American family. It can be a single-parent family.

Those mothers need help. And one way to do it is to get these deadbeat fathers to pay their obligations to these mothers – that will help strengthen the American family. And there's a whole bunch of other things that I can't click off in this short period of time.

LEHRER: All right, Mr. Perot, you have one minute.

PEROT: If I had to solve all the problems that face this country and I could be granted one wish as we started down the trail to rebuild the job base, the schools and so on and so forth, I would say a strong family unit in every home, where every child is loved, nurtured, and encouraged. A little child before they're 18 months learns to think well of himself or herself or poorly. They develop a positive or negative self- image. At a very early age they learn how to learn. If we have children who are not surrounded with love and affection – you see, I look at my grandchildren and wonder if they'll ever learn to walk because they're always in someone's arms. And I think, my gosh, wouldn't it be wonderful if every child had that love and support. But they don't. We will not be a great country unless we have a strong family unit in every home. And I think you can use the White House as a bully pulpit to stress the importance of these little children, particularly in their young and formative years, to mold these little precious pieces of clay so that they, too, can live rich full lives when they're grown.

D.3.4 Segment summary statistics across SITS and hand-coding approaches

I explore potential avenues that could make the SITS and hand-coded segments different in ways that allowed MTurk Workers to rate the SITS segments with greater cluster quality.

First, it is not that SITS segments were shorter and potentially easier to read and rate. Nor were SITS segments longer, potentially containing more information to make an accurate rating. I conduct a t-test of the length in words of the 100 hand-coded segments and 100 SITS segments used in the validation exercise. Hand-coded segments had a mean of 369 words and SITS segments had a mean of 420, which was an insignificant difference ($p = 0.365$).

Second, I assess the different length *within* pairs across the two approaches. I conduct a t-test of the absolute difference in length in words within the 50 hand-coded pairs of segments and the 50 SITS pairs of segments used in the validation exercise. Hand-coded pairs had an average difference of 424 words and SITS pairs had an average difference of 418 words, which was an insignificant difference between the approaches ($p = 0.938$). The results suggest that the significant result reported in the paper can not be attributed to difference in length between the approaches' segments or pairings.

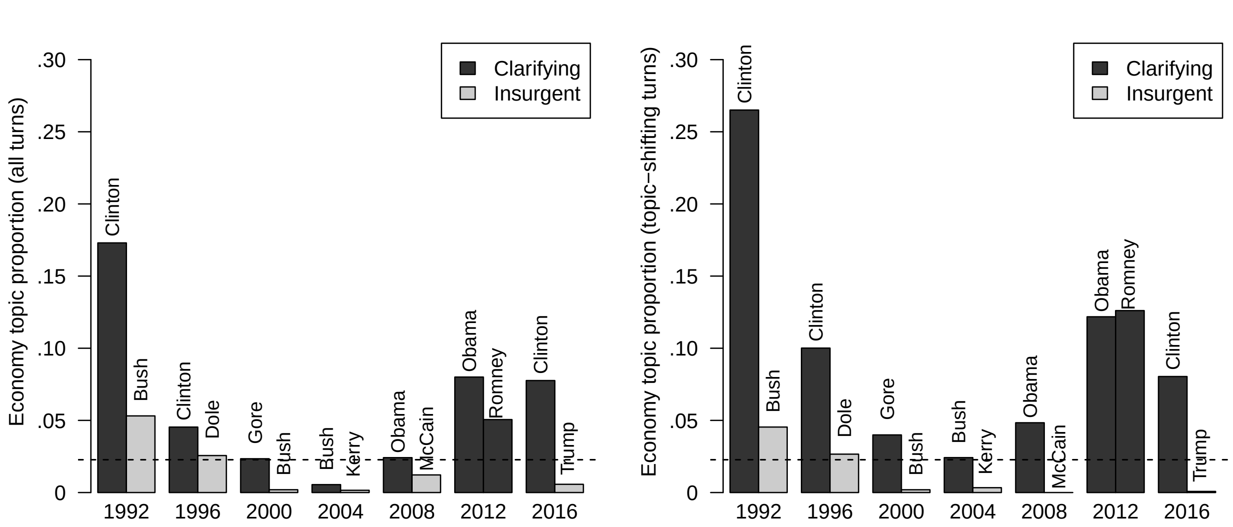
E PRESIDENTIAL DEBATES

E.1 Replication with alternative preprocessing

As discussed in Appendix B, `preText` result suggest I replicate results when preprocessing the text in two ways — once removing stopwords and once retaining stopwords. I include results for removing stopwords in the main body of the paper. Also, convergence and topic alignment results in Appendix E.2 also pertain to model results when removing stopwords. However, I replicate the substantive results reported in the paper when modeling the text without removing stopwords when preprocessing.

Figure 6 presents the economy topic proportions for clarifying and insurgent candidates when modeling the text without removing stopwords. We see that results are largely consistent with what is reported in the main body of the paper. The main inconsistency is Kerry's higher rate of shifting to the economy in Figure 6, which I did not expect would result from retaining stopwords in the corpus.

Figure 6: Clarifying candidates talk more about the economy



(a) All turns

(b) Topic-shifting turns

Note: Topic proportions for all of a candidate's speaking turns in (a) and all of the turns in which they shifted topic in (b). Dashed line represents if they talked about all topics equally. Clarifying candidates discuss the economy more than insurgent candidates, and engage in agenda setting to shift the course of the debate to the economy.

E.2 Convergence and topic alignment

Using the Gelman diagnostic (Gelman and Rubin 1992; Brooks and Gelman 1998) for the three estimated chains (see Appendix C), I find that $R_c < 1.2$ is reached for 91% of the 899 topic shift indicators, thus I am fairly confident convergence was reached.

I do not formally assess convergence of the topic assignment latent variables estimated by the model $(z_{d,t,n})$ for two reasons. First, there are $V \cdot K$, where V is the length of the corpus vocabulary, parameters to assess convergence. In this case, 944 words and 44 topics results in 41,536 parameters. Second, traditional convergence diagnostics are ill-equipped for categorical variables. So instead, I assess topic coherence and alignment across the models. Topic coherence, when considering the 10 most probable words for each topic, is consistent across models (Mimno et al. 2011). The total coherence for all 44 topics across the 3 models is -3637.841, -3547.431, and -3539.433, respectively. Model three has the largest coherence metric. Next, I assess topic alignment. I consider all permutations of the three models. For each topic in the first model of consideration (the reference model), I choose the topic from second model of consideration (the candidate model) that yields the minimum inner product between the K by V topic matrices. We see that the best alignment occurs between models two and three (and vice versa). Likewise, the median number of shared words is 7 (out of 10) between the aligned topics of models two and three.

Candidate	Reference	Average $L1$	Median number of shared top words
1	2	0.730	7
1	3	0.780	6
2	1	0.714	7
2	3	0.661	7
3	1	0.761	6
3	2	0.687	7

E.3 Topics

Because model three has the best performance across the metrics, I present FREX top words from this model in Table 1. I also present labels for the topics. Labels were chosen from (1) listening to audio recordings and multiple readings each of deliberation's transcript, (2) consulting speaking turns and segments highly associated with each topic, and (3) consulting the FREX top words.

Table 1: Presidential debate topics

Label	Top words
Enforcing time limits	sir, minut, pleas, respond, goe, sorri, ask, statement
North Korea	nuclear, iran, korea, north, weapon, sanction, threat, tabl
Energy	energi, coal, gas, oil, clean, land, technolog, wind
Affirmative action	administr, action, vice, discrimin, upon, subject, feel, strong
Bipartisanship	republican, democrat, togeth, commiss, polit, done, bipartisan, senat
Campaign promises	campaign, word, apolog, anyth, fact, never, foreign, week
Healthcare	health, insur, care, cost, cover, afford, obamacar, premium
Immigration	border, immigr, illeg, citizen, law, open, employ, wait
Iraq War	saddam, hussein, osama, bin, laden, destruct, weapon, mass
Government spending	spend, billion, defens, propos, dollar, budget, prioriti, cut
Economy	invest, econom, economi, growth, grow, creat, unemploy, trickl
Answer transition	point, mention, import, obvious, face, critic, situat, hand
Trade	trade, agreement, export, market, free, sell, japan, product
Federal deficit	debt, trillion, pay, share, grow, money, segment, build
College education	colleg, young, opportun, kid, abl, teacher, loan, grant
Government operations	congress, night, littl, control, program, industri, interest, run
Government budget	deficit, deduct, balanc, budget, cut, trillion, revenu, add
Social welfare	welfar, budget, capit, veto, educ, gain, child, centuri
Family	love, experi, children, faith, daughter, man, part, person
Terrorism	terrorist, train, free, safe, allianc, safer, world, iraqi
Role of government	feder, texa, govern, size, role, differ, vice, prioriti
Small business	busi, small, rate, percent, credit, tax, high, lawyer
ISIS	isi, syria, attack, muslim, fli, arm, zone, fight
Taxes	tax, incom, class, middl, relief, pay, rais, corpor
Problems	bad, mani, disast, great, lie, ever, actual, deal
Defense	militari, world, nation, forc, peac, mission, around, troop
Financial crises	bank, financi, crisi, mortgag, street, home, regul, wall
Outsourcing work overseas	china, compani, oversea, industri, competit, invest, manufactur, worker
Education	school, teacher, parent, educ, public, children, choic, student
Hope	thank, togeth, hope, trust, promis, futur, elect, opportun
Russia	russian, putin, russia, nato, democraci, old, clear, cold
Women's rights	women, valu, famili, men, rais, woman, wage, leav
War in Afghanistan	troop, afghanistan, iraq, strategi, qaida, succeed, general, war
Abortion	abort, life, woman, aid, research, law, allow, ban
Race & policing	african, citi, inner, communiti, racial, race, terribl, realli
Medicare	senior, medicar, prescript, drug, lower, current, compani, price
Political experience	record, vote, oppos, senat, financ, fought, parti, reform
Gun control	gun, polic, crime, assault, law, ban, check, death
Mistakes	mistak, serious, tri, wrong, mayb, charact, seem, disagre
Supreme Court	court, judg, constitut, suprem, justic, amend, appoint, decis
Funding education	drug, live, fight, better, histori, children, respons, centuri
Social security	social, secur, trust, surplus, fund, promis, benefit, retir
Middle East	east, israel, middl, leadership, region, friend, peac, cours

Debate salutations	presidenti, candid, debat, univers, welcom, tonight, commiss, nomine

E.4 Topic-shifting by party

In Figure 7 I present topic proportions by party, aggregating across the entire corpus of 20 debates, in turns identified as shifting the topic. Asterisks indicate significant differences ($p < .05$) from a difference in proportions text. The dashed line represents if candidates talked about all topics equally. First we see some familiar patterns. Democrats shift to topics like health care, social welfare, and energy more often than other topics, and they shifted to these topics more often than Republicans did. Likewise, Republicans shifted to topics like the role of government, small businesses, and race & policing (or "law & order") more often than Democrats.

Because I am aggregating attention to these topics from elections that span 25 years, some topics that emerge later during the time period, such as ISIS, are shifted to less frequently. Therefore, it may be useful if aggregating across time, to compare how often topics are shifted to against how much they are discussed in the corpus at-large. ISIS may not be discussed often in absolute terms, but in the elections that it does appear, it may take a prominent role on a candidate's agenda.

Something else to note is the "Problems" topic. This is almost exclusively a topic used by Trump (and the topic Trump almost exclusively uses) to draw attention to negative attributes of the current state of politics. Again, this presents an instance where it may be useful if aggregating across time, to compare how often topics are shifted to against how much they are discussed in the corpus at-large.

E.5 Example topic changes

Table 2 presents a tw examples of estimated of topic shifts and the highest probability topic of the new segment. The first example, referenced in the main body of the article, occurs at the end of the first 2016 general election debate. This example is a particularly contentious exchange, and we see an estimated topic shift when Clinton shifts the topic to Trump's history of insulting women. In fact, Trump starts engaging in Clinton's shifted-to topic, evidence of Clinton's higher agenda-setting ability relative to Trump. The second example is an instance where candidates stay *on* topic in their answers to a question posed by an audience member in the second debate, and the model estimates no topic changes occurred. After their initial answers, the candidates engage in a back-and-forth on Obamacare until the next question is asked by an audience member, all of which is considered one segment of the debate, estimated to mainly be on the healthcare topic.

Table 2: Example topic changes in 2016 presidential debate

Example 1– shifting topic

⋮

Clinton: We are at—we are at the final question.

Clinton: Well, one thing. One thing, Lester.

Moderator: Very quickly, because we're at the final question now.

Women's rights (women, valu, famili, men, rais, woman, wage, leav)

Clinton: You know, he tried to switch from looks to stamina. But this is a man who has called women pigs, slobs and dogs, and someone who has said pregnancy is an inconvenience to employers, who has said

Trump: I never said that.

Clinton: ...women don't deserve equal pay unless they do as good a job as men.

Trump: I didn't say that.

⋮

Example 2 – staying on topic

⋮

Moderator: Ken Karpowicz has a question.

Trump: It’s nice to—one on three.

Health care (health, insur, care, cost, cover, afford, obamacar, premium))

Audience Question: Thank you. Affordable Care Act, known as Obamacare, it is not affordable. Premiums have gone up. Deductibles have gone up. Copays have gone up. Prescriptions have gone up. And the coverage has gone down. What will you do to bring the cost down and make coverage better?

Moderator: That first one goes to Secretary Clinton, because you started out the last one to the audience.

Clinton: If he wants to start, he can start. No, go ahead, Donald.

Trump: No, Im a gentleman, Hillary. Go ahead.

Moderator: Secretary Clinton?

Clinton: Well, I think Donald was about to say hes going to solve it by repealing it and getting rid of the Affordable Care Act. And Im going to fix it, because I agree with you. Premiums have gotten too high. Copays, deductibles, prescription drug costs, and Ive laid out a series of actions that we can take to try to get those costs down. But heres what I don’t want people to forget when were talking about reining in the costs, which has to be the highest priority of the next president, when the Affordable Care Act passed, it wasn’t just that 20 million people got insurance who didn’t have it before. But that in and of itself was a good thing. I meet these people all the time, and they tell me what a difference having that insurance meant to them and their families. But everybody else, the 170 million of us who get health insurance through our employers got big benefits. Number one, insurance companies cant deny you coverage because of a pre-existing condition. Number two, no lifetime limits, which is a big deal if you have serious health problems. Number three, women cant be charged more than men for our health insurance, which is the way it used to be before the Affordable Care Act. Number four, if you’re under 26, and your parents have a policy, you can be on that policy until the age of 26, something that didn’t happen before. So I want very much to save what works and is good about the Affordable Care Act. But we’ve got to get costs down. We’ve got to provide some additional help to small businesses so that they can afford to provide health insurance. But if we repeal it, as Donald has proposed, and start over again, all of those benefits I just mentioned are lost to everybody, not just people who get their health insurance on the exchange. And then we would have to start all over again. Right now, we are at 90 percent health insurance coverage. Thats the highest we’ve ever been in our country.

Moderator: Secretary Clinton, your time is up.

Clinton: So I want us to get to 100 percent, but get costs down and keep quality up.

Moderator: Mr. Trump you have two minutes.

Trump: It is such a great question and its maybe the question I get almost more than anything else, outside of defense. Obamacare is a disaster. You know it. We all know it. Its going up at numbers that nobody’s ever seen worldwide. Nobody’s ever seen numbers like this for health care. Its only getting worse. In 17, it implodes by itself. Their method of fixing it is to go back and ask Congress for more money, more and more money. We have right now almost $20 trillion in debt. Obamacare will never work. Its very bad, very bad health insurance. Far too expensive. And not only expensive for the person that has it, unbelievably expensive for our country. Its going to be one of the biggest line items very shortly. We have to repeal it and replace it with something absolutely much less expensive and something that works, where your plan can actually be tailored. We have to get rid of the lines around the state, artificial lines, where we stop insurance companies from coming in and competing, because they want and President Obama and whoever was working on it they want to leave those lines, because that gives the insurance companies essentially monopolies. We want competition. You will have the finest health care plan there is. She wants to go to a single- payer plan, which would be a disaster, somewhat similar to Canada. And if you haven’t noticed the Canadians, when they need a big operation, when something happens, they come into the United States in many cases because their system is so slow. Its catastrophic in certain ways. But she wants to go to single payer, which means the government basically rules everything. Hillary Clinton has been after this for years. Obamacare was the first step. Obamacare is a total disaster. And not only are your rates going up by numbers that nobody’s ever believed, but your deductibles are going up, so that unless you get hit by a truck, you’re never going to be able to use it.

⋮

F IN-PERSON DELIBERATIONS

F.1 Correlation matrix details

Table 3 presents Pearson's r correlation coefficients for the corresponding visualization in the paper. Agenda setting is negatively correlation with how often a participant interrupts others, but there is no evidence of a correlation between the other count-based measures based on amount of participation. Taken together, these results replicate the finding that agenda setting is not achieved by out-talking or interrupting others.

Table 3: Correlations between agenda setting and participation measures

	Agenda setting	Proportion of comments	Proportion of talk time	Proportion of interruptions	Composite measure
Agenda setting	1.00	-0.07	-0.03	-0.33*	-0.19
Proportion of comments		1.00	0.88*	0.21	0.93*
Proportion of talk time			1.00	-0.04	0.81*
Proportion of interruptions				1.00	0.52*
Composite measure					1.00

F.2 Convergence and topic alignment

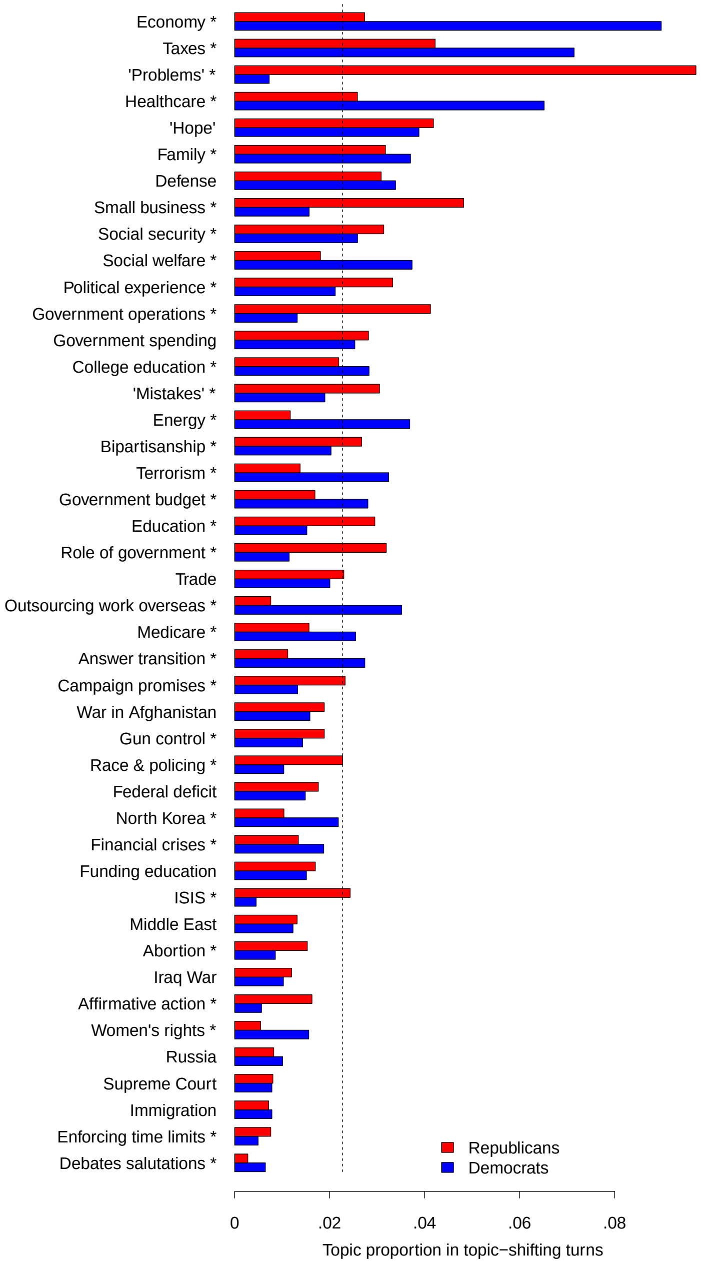
I outline my approach to assessing convergence and topic alignment in Appendix E.2. For this application, I find that $R_c < 1.2$ is reached for 83% of the 899 topic shift indicators, thus I am fairly confident convergence was reached. The total coherence for all 18 topics across the 3 models is -1451.491, -1499.628, and -1448.98, respectively. Model three has the largest coherence metric. Lastly, we see that the best alignment occurs between models one and three (and vice versa). Likewise, the median number of shared words is 6.5 (out of 10) between the aligned topics of models one and two.

Candidate	Reference	Average $L1$	Median number of shared top words
1	2	0.802	6
1	3	0.567	6.5
2	1	0.828	6
2	3	0.693	6
3	1	0.576	6.5
3	2	0.780	6

F.3 Topics

Because model three has the best performance across the metrics, I present FREX top words from this model. I also present labels for the topics in Table 4. Labels were chosen from (1) listening to audio recordings and multiple readings each of deliberation's transcript, (2) consulting speaking turns and segments highly associated with each topic, and (3) consulting the FREX top words.

Figure 7: Topic proportions in topic-shifting turns, by party



F.4 Replication with alternative preprocessing

As discussed in Appendix B, `preText` result suggest I replicate results when preprocessing the text in two ways — once removing stopwords and once retaining stopwords. I include results for removing stopwords in the main body of the paper. Convergence and topic alignment results in Appendix G.3 also pertain to model results when removing stopwords. However, I replicate the substantive results reported in the paper when modeling the text without removing stopwords when preprocessing. Table 5 presents Pearson's r correlations which replicate the main results when retaining stopwords during preprocessing. Table 6 replicates the sign and significance of the regression coefficient on agenda setting.

Table 4: In-person deliberation topics

Label	Top words
Enforcement at testing center	test, center, shave, hard, go, turn, away, take
Pants-length rules	short, knee, shoe, someth, super, wear, mean, girl
Judgment amongst students	differ, realli, also, kindof, judg, feel, dress, care
Women's guidelines	women, leg, fit, wear, sleeveless, guy, garment, cloth
Coordinating proposals	say, good, want, write, idea, can, option, yeah
Personal benefits of DGS	rule, push, follow, job, interview, tri, want, less
Including photo examples of DGS	photo, pictur, exampl, cloth, includ, groom, qualiti, dress
Deliberation guidelines	vote, group, research, inform, confidenti, pass, pleas, ballot
Lack of consistent enforcement	code, enforc, consist, honor, teacher, syllabus, part, run
Facial hair on campus	came, bit, littl, general, ever, school, better, shaven
Facial hair rule	facial, hair, trim, beard, mustach, long, men, must
Five minute warning	minut, five, continu, propos, discuss, may, begin, readi
Ear-piercings rule	pierc, ear, lds, see, tell, accord, mormon, prophet
Professional-appearing facial hair	beard, grow, look, mayb, thing, profession, someon, guy
Hairstyle rules	color, extrem, avoid, bun, style, line, hairstyl, purpl
Purpose of DGS (cultural)	reason, cultur, us, bodi, back, know, set, someth
Polling students	univers, student, opinion, everi, poll, structur, year, major
Purpose of DGS (spiritual)	church, reflect, purpos, guess, standard, youth, valu, non

Table 5: Replication of: Agenda setting does not correlate with quantity of participation

	Agenda setting	Proportion of comments	Proportion of talk time	Proportion of interruptions	Composite measure
Agenda setting	1.00	-0.10	-0.01	-0.34*	-0.20
Proportion of comments		1.00	0.88*	0.21	0.93*
Proportion of talk time			1.00	-0.04	0.81*
Proportion of interruptions				1.00	0.52*
Composite measure					1.00

G ONLINE DISCUSSIONS

G.1 MTurk task

In this case, the task was available to any worker that was 18 years or older, was located in the United States, held a 99 percent or higher task approval rating, and had completed 2000 or more tasks. Participants were paid the agreed compensation within the MTurk platform within hours of completing the task.

G.2 Information on charities provided to participants

Respondents were provided the following information about each charity via hover boxes on Qualtrics when choosing which charity they'd prefer the researchers donate $1.00 to. The information was also available to participants when discussing the charities with their assigned partner.

Information is provided about each charity from Charity Navigator. Charity Navigator rates charities by evaluating Financial Health and Accountability & Transparency using financial information from each charity's informational tax return and website. Their ratings "show donors how efficiently a charity will use their support, how well it has sustained its programs and services over time, and their level of commitment to accountability and transparency."

Hover your mouse over each charity to read its mission statement and view its financial health and accountability & transparency scores. Clink on the provided link for additional information.

Table 6: Replication of: Agenda setters more likely to shape deliberation outcome

	Dependent variable: Introduced group proposal
Agenda setting	10.86^{*}
	(4.78)
Strong DGS attitudes	.867
	(1.03)
Group indicators	✓
Constant	-3.67^{*}
	(1.48)
Observations	40
AIC	57.47

Note: *p<0.05. Coefficient from logistic regression with clustered standard errors at the discussion-group level in parentheses. Dependent variable is introducing one of up to two ideas included as a group policy proposal ($y = 1$) or not ($y = 0$).

- American Red Cross—Since its founding in 1881 by visionary leader Clara Barton, the American Red Cross has been the nation's premier emergency response organization. We bring shelter, food and comfort to those affected by disasters, large and small. We collect lifesaving donated blood and supply it to patients in need. We provide support to our men and women in military bases around the world, and to the families they leave behind. We train communities in CPR, first aid and other skills that save lives. And we assist our neighbors abroad with critical disaster response, preparedness and disease prevention efforts. We are able to do all this by mobilizing the power of volunteers and the generosity of donors.
  Financial Health score: $77.50/100$
  Accountability & Transparency score: $100/100$
  Read more here.

- ALSAC - St. Jude Children's Research Hospital—ALSAC (American Lebanese Syrian Associated Charities) was founded in 1957 and exists for the sole purpose of raising funds to support the operating and maintenance of St. Jude Children's Research Hospital. The mission of St. Jude Children's Research Hospital is to find cures for children with cancer and other catastrophic diseases through research and treatment. It is supported primarily by donations raised by ALSAC. Research efforts are directed at understanding the molecular, genetic and chemical bases of catastrophic diseases in children; identifying cures for such diseases; and promoting their prevention. Research is focused specifically on cancers, some acquired and inherited immunodeficiencies, sickle cell disease, infectious diseases and genetic disorders.
  Financial Health score: $87.33/100$
  Accountability & Transparency score: $100/100$
  Read more here.

- Doctors Without Borders, USA—Doctors Without Borders, USA (DWB-USA) was founded in 1990 in New York City to raise funds, create awareness, recruit field staff, and advocate with the United Nations and US government on humanitarian concerns. Doctors Without Borders/Mdecins Sans Frontires (MSF) is an international medical humanitarian organization that provides aid in nearly 60 countries to people whose survival is threatened by violence, neglect, or catastrophe, primarily due to armed conflict, epidemics, malnutrition, exclusion from health care, or natural disasters.
  Financial Health score: $97.50/100$
  Accountability & Transparency score: $97.50/100$
  Read more here.

- UNICEF USA—The United Nations Children's Fund (UNICEF) works in more than 190 countries and territories to put children first. UNICEF has helped save more children's lives than any other humanitarian organization, by providing health care and immunizations, clean water and sanitation, nutrition, education, emergency relief and more. UNICEF USA supports UNICEF's work through fundraising, advocacy and education in the United States.

Together, we are working toward the day when no children die from preventable causes and every child has a safe and healthy childhood.
Financial Health score: 75.00/100
Accountability & Transparency score: 97.00/100
Read more here.

- American Heart Association—The American Heart Association is the nation's oldest and largest voluntary organization dedicated to fighting heart disease and stroke. To improve the lives of all Americans, we provide public health education in a variety of ways. We're the nation's leader in CPR education training. We help people understand the importance of healthy lifestyle choices. We provide science-based treatment guidelines to healthcare professionals to help them provide quality care to their patients. We educate lawmakers, policymakers and the public as we advocate for changes to protect and improve the health of our communities. We have funded more than $3.8 billion in heart disease and stroke research, more than any organization outside the federal government. With your help, we are working toward improving the cardiovascular health of all Americans by 20 percent, and reducing deaths from cardiovascular diseases and stroke by 20 percent, all by the year 2020.
  Financial Health score: 88.62/100
  Accountability & Transparency score: 96.00/100
  Read more here.

G.3 Convergence and topic alignment

I outline my approach to assessing convergence and topic alignment in Appendix E.2. For this application, I find that $R_c < 1.2$ is reached for 93% of the 1516 topic shift indicators, thus I am fairly confident convergence was reached. The total coherence for all 17 topics across the 3 models is -1714.849, -1726.524, and -1771.158, respectively. Model one has the largest coherence metric. Lastly, we see that the best alignment occurs between models one and two (and vice versa). However, the median number of shared words is 9 (out of 10) between the aligned topics of models one and three.

Candidate	Reference	Average $L1$	Median number of shared top words
1	2	0.385	8
1	3	0.420	9
2	1	0.391	8
2	3	0.485	8
3	1	0.439	9
3	2	0.487	8

G.4 Topics

Because model one has the best performance across the metrics, I present FREX top words from this model. I also present labels for the topics in Table 7. Labels were chosen from (1) multiple readings each of deliberation's transcript, (2) consulting speaking turns and segments highly associated with each topic, and (3) consulting the FREX top words.

Table 7: Online discussion topics

Label	Top words
St. Jude's (STJ)	research, child, kid, cancer, sick, children, everyon, cure
American Heart Association (AHA)	heart, aha, friend, famili, die, member, old, cancer
Terms of deliberation	decis, unanim, decid, bonus, go, let, make, sorri
UNICEF	unicef, kid, pick, true, point, children, fact, around

Saying "hello"	chose, hello, hi, choos, dwb, stj, prefer, hey
Deliberating	way, either, okay, fine, say, feel, appreci, go
Need for donations	peopl, need, import, differ, help, money, goe, servic
Charities' missions	medic, provid, care, treat, countri, access, healthcar, without
Charity Navigator scores	score, financi, highest, transpar, 100, high, health, account
Time remaining	minut, 10, still, guess, want, full, left, wait
Saying "goodbye"	nice, thank, day, talk, bye, chat, sound, _smiley_
Reaching disagreement	much, know, seem, pretti, dont, like, heard, feel
Reaching agreement	money, definit, thing, open, agre, total, mayb, wonder
St. Jude's commercials	yea, commerci, watch, year, now, lol, last, bad
Deliberating	will, switch, choic, yeah, alreadi, red, second_choic, list
Doctors Without Borders (DWB)	volunt, doctor, understand, children, admir, lean, complet, heard
American Red Cross (ARC)	disast, hurrican, arc, natur, help, experi, affect, tornado

References

Brooks, Stephen P and Andrew Gelman. 1998. "General Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational and Graphical Statistics* 7(4):434–455.

Denny, Matthew J and Arthur Spirling. 2018. "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It." *Political Analysis* 26(2):168–189.

Gelman, Andrew and Donald B Rubin. 1992. "Inference from Iterative Simulation using Multiple Sequences." *Statistical Science* 7(4):457–472.

Griffiths, Thomas L and Mark Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National academy of Sciences* 101(Suppl. 1):5228–5235.

Mimno, David, Hanna M Wallach, Edmund Talley, Miriam Leenders and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics pp. 262–272.

Mozer, Reagan, Luke Miratrix, Aaron Russell Kaufman and L Jason Anastasopoulos. Forthcoming. "Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality." *Political Analysis* .

Nguyen, Viet-An. 2014. "Speaker Identity for Topic Segmentation (SITS)." GitHub repository. https://github.com/vietansegan/sits. Last accessed September 25, 2019.

Nguyen, Viet-An, Jordan Boyd-Graber, Philip Resnik, Deborah A Cai, Jennifer E Midberry and Yuanxin Wang. 2014. "Modeling Topic Control to Detect Influence in Conversations Using Nonparametric Topic Models." *Machine Learning* 95(3):381–421.

Nikita, Murzintcev. 2020. *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters*. R package version 1.0.2.
URL: *https://CRAN.R-project.org/package=ldatuning*

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4):1064–1082.