Measuring Agenda Setting Power in Interactive Political Communications*

Erin Rossiter

Department of Political Science

Washington University in St. Louis

November 21, 2019

ABSTRACT

How can we determine the relative power of actors in interactive political communications, such as debates, deliberations, and discussions? The literature provides one answer to this question: actors can exert power by *agenda setting*, or shifting the conversational agenda to preferred topics. However, political scientists currently lack a method for identifying and accurately measuring this core concept. In this article, I measure agenda setting power with a topic model that attributes topic changes throughout the interaction to actors (Nguyen et al. 2014). Simulations show that this measure performs well even with rapid dialogue typical of interactions and that existing text as data approaches in political science do not accomplish this task. I validate this measure across three empirical applications: the 2016 US presidential debates, in-person deliberations, and a novel online deliberation study. These applications also demonstrate the importance of studying agenda setting power as a key element of political interactions.

*Valuable feedback for this project was provided by Political Data Science Lab members at Washington University in St. Louis. I thank participants at VIM 2019, New Faces in Political Methodology XI, Text as Data 2018, APSA 2018, PolMeth 2018, and SPSA 2018 for helpful comments. I am grateful to Christopher Karpowitz and Hans Hassell for generously sharing data. Funding for this project was provided by the National Science Foundation (SES-1558907). I appreciate any comments or suggestions. Please do not cite or distribute without permission.

1 INTRODUCTION

Who holds the upper-hand, the control, or the *power* in a political exchange is fundamental to the study of political discourse but remains an elusive concept to quantify. While observing the exercise of power is more straightforward in the formal political arena (e.g., a president vetoes a bill), we lack a systematic way to study power in interactive communications among actors. Yet measuring power in these settings is important because debates, deliberations, and discussions precede any formal display of power we observe. In this article, I draw on research from computer science to measure one form of power in formal and informal political interactions—agenda setting. I show in my applications that actors seek to set the agenda by controlling what is (and what is not) discussed. Further, I show that this method outperforms standard approaches used in political science to analyze behavior in interactive communications.

An overwhelming amount of politics occurs through interactive political communications. Presidential debates, Congressional committee hearings, and Supreme Court oral arguments are all examples of formalized political interactions that are embedded in the framework of American government to facilitate decision-making. Research shows these interactions play a role in the outcomes we subsequently observe, such as the ability of a lawyer's oral argument to sway Supreme Court votes (Johnson, Wahlbeck and Spriggs 2006). Beyond these formal settings, informal interactions are also ubiquitous in politics as they are fundamental to the work of politicians, lobbyists, bureaucrats, and more. Additionally, whether talking around the dinner table or watching panels of pundits on television, citizens largely experience politics via interactions. Research shows that engaging in political conversation can influence subsequent behaviors like vote choice (Beck et al. 2002).

Despite the significance of interactions in the political sphere, political methodology has so far overlooked some of their unique features. Interactions—unlike other political communications like press releases, speeches, or party manifestos—are a social game. While actors engage in conversation, a complex social exercise is underway as actors negotiate who has the floor and what is being discussed in real-time. Importantly, the social nature of interactions invites the use of agenda

setting as an effective way to exercise power. Actors can shape how the interaction unfolds—what issues or topics are and are not discussed—through attempts to shift the conversational agenda to preferred topics. Whether public, confrontational debates or private, thoughtful deliberations, actors have an incentive to optimize the interaction. This incentive stems from the two-fold impact of successful agenda setting within an interaction. Agenda setting leads to both an immediate control over what is discussed, and because of this, agenda setting may also impact subsequent outcomes stemming from the interaction.

Take, for example, an interactive setting in which the fight over the agenda is particularly evident—United States presidential debates. Candidates seek to set the agenda, or to shift the debate toward topics they "own" or to those that may harm their opponent (Petrocik 1996; Boydstun, Glazier and Phillips 2013). The following lines from the first 2016 general election presidential debate between Donald Trump and Hillary Clinton, with moderator Lester Holt, demonstrate Clinton shifting the agenda to her preferred discussion topic.[1]

> *Holt:* We are at—we are at the final question.
>
> *Clinton:* Well, one thing. One thing, Lester.
>
> *Holt:* Very quickly, because we're at the final question now.
>
> *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...

Holt, as the moderator, tried to introduce his final debate question. Yet Clinton, in real-time, overpowered his efforts and successfully steered the final minutes of the debate toward an issue on *her* agenda—Trump's history of degrading women. Clinton's skill at setting the agenda not only affected what was discussed during the debate, but it also influenced subsequent media converge. As discussed below, news outlets reported that this exchange was a memorable moment from the first debate (Ross 2016; Mason 2016).

Although it is easy to see why measuring agenda setting power in these settings is important, standard approaches in political science offer no way to quantify Clinton's power over what was

[1] Transcript from Commission on Presidential Debates (2016).

discussed during the debate. Broadly, political scientists lack a systematic way to measure the relative power of actors in interactive settings. To be sure, political scientists have worked to quantify behaviors in political interactions. A common approach is to count easily observable and quantifiable behaviors, such as the number of words spoken, the number of questions asked, or the number of interruptions made by a participant. Such count-based measures are popular across a variety of political contexts, such as citizen deliberations, (e.g., Karpowitz, Mendelberg and Shaker 2012) legislative committee hearings (e.g., Kathlene 1994), and Supreme Court oral arguments (e.g., Epstein, Landes and Posner 2010). Yet, while count-based measures are useful for studying *observable quantities*, such as how much an actors participates or who they tend to interrupt, these measures fall short when the goal is to measure any *latent quantities* in an interaction, such as power.

In this article, I build upon these previous efforts in political science to quantify important behaviors within interactions, and I focus my efforts on measuring the latent agenda setting power of actors. Specifically, I leverage the text of interactions as data, and I use the Speaker Identity for Topic Segmentation (SITS) model from the computer science topic segmentation literature (Nguyen et al. 2014; Nguyen, Boyd-Graber and Resnik 2012). SITS extends Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan 2003), a topic model used widely in political science, to simultaneously estimate *where shifts in the agenda occur* within an interaction and *each actor's agenda setting power* (Nguyen 2015).

I proceed by first considering how agenda setting in political interactions fits within the broader context of theories of political power. I argue that agenda setting power in interactions corresponds conceptually to previously identified dimensions of power. I then outline an approach used in computer science to identify agenda setting in multi-speaker settings (Nguyen 2015). I then present two sets of simulation results. First, I show that the model performs well even with rapid dialogue which typifies many political interactions. Second, I demonstrate that existing text as data approaches in political science are not well-suited for the task of measuring the agenda setting power of actors in an interactive setting.

Then across three applications, I show that the measure captures the theoretical construct of power and that agenda setting correlates with important outcomes of political interactions. I first measure the agenda setting power of candidates in electoral debates.[2] I find candidates' agenda setting power aligns with media accounts of their performance, suggesting this measure provides a valid way to capture agenda setting power in this setting. Next, I assess how agenda setting relates to the common approach in political science of using count-based measures of participation to quantify speaker behavior in interactive settings. Using deliberation texts, I find that a deliberator's agenda setting power correlates with her attitude change, while count-based measures of participation do not. Lastly, I perform a direct test of my central claim that agenda setting is a form of power. In a novel online deliberation study, I show that an individual's agenda setting power correlates with 'winning' or achieving a desired outcome. I conclude with a discussion of the usefulness of studying agenda setting power as an important element of political interactions, and I provide suggestions for further areas of inquiry enabled by this measure of power.

2 AGENDA SETTING AS A FORM OF POWER

Robert Dahl's 1957 essay on power sought to formally define the concept so that political scientists could operationalize power and study the relative power of actors (Dahl 1957). While his theory focuses on the behavior of actors in, and the outcome of, a decision-making situation, Dahl's operationalization of power was admittedly limited due to what researchers could and could not observe and measure at that time. In addressing his research on relative power of senators, Dahl writes, "Faced with this apparently insuperable obstacle, it was necessary to adopt a rather drastic alternative, namely to take the recorded roll-call vote of a Senator as an indication of his position and activities *prior to* the roll-call" (Dahl 1957, p. 210). Dahl acknowledges that what precedes formal, observable decision-making stages of politics is important yet not taken into account in the empirical stages of his work on power.

The fight for and exercise of power extends beyond observable, concrete decisions. Actors do not make their decisions in a vacuum, but instead deliberate and discuss formally and informally

[2] I build upon the work of Nguyen (2015) who measures agenda setting in presidential debates.

before taking decisive action. To this end, Bachrach and Baratz (1962) introduce a second dimension of power. They call this dimension "nondecision-making power," as it precedes the formal decision-making stage of politics and is "invisible" if one consults only the outcomes of a decision.

If this form of power is unobservable when studying the outcomes of a decision-making situation, how does it manifest? Bachrach and Baratz argue "power may be, and often is, exercised by confining the scope of decision-making" (Bachrach and Baratz 1962, p. 948). Indeed, Lukes arrives at a similar dimension of power which he terms "agenda setting" power, as power over the decision-making stage can occur by shaping the agenda *before* any votes, for example, are cast (Lukes 1974). Building on these views, I argue *interactions* invite the use of agenda setting power. In an interaction, agenda setting power manifests as actors seek to shape the scope of the conversation by gaining the floor and maintaining attention on their preferred issues.

Because agenda setting shapes the set of important issues or topics that result from an interaction, agenda setting within political interactions has the potential to influence formal, decision-making outcomes. In other words, debates, deliberations, and discussions provide actors the opportunity to alter the scope of a given conflict (Schattschneider 1975). For example, successfully setting the agenda might result in keeping certain issues off the table, obviating the risk of the issue rising to a vote. Or, conversely, effective agenda setting might raise the status of an otherwise overlooked issue, altering what occurs at later decision-making stages. Because of agenda setting's impact on what is discussed in the immediate interaction and its potential impact on shaping subsequent political outcomes, actors have an incentive to exercise agenda setting power.

To be sure, a great deal of research has been done regarding how interactions unfold in different political contexts and how interactions impact the attitudes of participants and political outcomes. Consider the literature on one type of political interaction—deliberation. Scholars have empirically investigated how participation is influenced by the deliberation's decision-making rule (e.g., Karpowitz, Mendelberg and Shaker 2012) and gender diversity (Kathlene 1994) of participants. Moreover, scholars have studied deliberation's impact on outcomes such as public opinion (e.g., Barabas 2004; Druckman and Nelson 2003).

These studies represent a broader practice of studying what is observable—whether it be behavior within interactions or the outcomes stemming from interactions. However, studying observable behavior does not allow us to understand the latent power exercised within an interaction or its downstream consequences. But if we take the conceptualization of power offered by Bachrach, Baratz, and Lukes seriously, then the agenda setting power of actors surely impacts how a political interaction unfolds and how it affects subsequent outcomes. Thus, I argue it remains an important task to understand agenda setting power within a political interaction, including who is able to gain such power, how they do so during the course of an interaction, and to what end.

2.1. *Possible approaches to measuring power in interactions in political science*

Political scientists currently lack a method for identifying and accurately measuring the core concept of power in interactions. Current tools used in political science for analyzing interactions are not suited for measuring latent behaviors of actors, hindering our ability to empirically investigate power dynamics in such settings.

One methods political scientists employ to learn about interactions is to field surveys about individuals' discursive habits (e.g., Huckfeldt, Johnson and Sprague 2004; Mutz 2006). Along these lines, one might consider fielding survey items to get opinions on one's own or others' agenda setting abilities. Yet, survey questions rely on self-reported behavior and perceptions, both tainted with potential bias (e.g., Prior 2009), and overlook the interactions themselves as rich sources of data.

Political scientists also used hand-coding methods to measures behaviors within interactions, so it is possible to hand code how effective each actor is at changing the topic to measure their agenda setting ability (e.g., Boydstun, Glazier and Phillips 2013; Boydstun, Glazier and Pietryka 2013). However, hand coding, while an intuitive and adaptable measurement strategy, has significant weaknesses. First, recruiting, adequately training, and compensating the work of research assistants can be prohibitively time-consuming and costly. Second, research shows that even high quality coders provide estimates that are unreliable (Mikhaylov, Laver and Benoit 2012).

Quantitative analysis of the text of interactions is a more popular approach. Scholars often

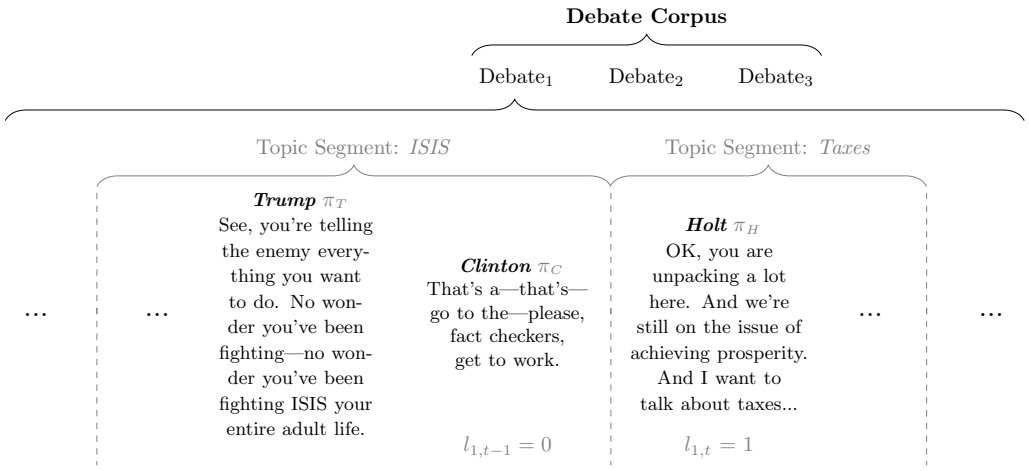
count observable and quantifiable behaviors such as the number of turns, interruptions, or words spoken by participants (e.g., Kathlene 1994; Epstein, Landes and Posner 2010; Karpowitz, Mendelberg and Shaker 2012). While this approach measures *quantity* of participation, itself an important concept in the study of representation in group deliberations, count-based measures are limited when the goal is to assess any sort of latent *quality* of the participation of an actor.

To be sure, automated text analysis has been applied to measure the concept of agenda setting (Quinn et al. 2010; Eggers and Spirling 2016). Quinn et al. (2010) conceptualize the agenda as what issues are broadly gaining attention in the political arena and which are not. As a macro-level measure of the agenda, it does not measure the micro-level agenda setting behavior of actors throughout the course of their everyday political interactions. Eggers and Spirling (2016) do indeed propose a measure of the agenda setting abilities of political actors. However, they conceptualize the agenda as the relative importance placed on issues over over months and years. Thus, even this measure is not a good fit for the study of agenda setting power in interactions, as interactions require the *immediate* negotiation of power as a communication unfolds.

2.2. *Proposed approach to measuring power in interactions*

In this section, I outline a framework for measuring the agenda setting power of actors in political interactions. Since an interactive political communication is a verbal or written exchange, I propose using the text of interactions (i.e., transcripts) as data. Above, I discuss that how political actors will seek to shift the topic of conversation to preferable or advantageous topics (and, in turn, avoid unpreferred or disadvantageous topics). Therefore, I apply a topic model to discover ***the latent topics discussed*** in the interaction. Knowing what is discussed is important only insofar as I can determine ***where shifts in agenda occur***. Thus, the topic model also identifies at what points throughout the course of the interaction the topic changes. Moreover, finding where shifts in topic occur is important because agenda setting power manifests as actors seek to shape the conversational agenda. Therefore, the topic model also simultaneously measures agenda setting power by identifying the ***extent to which each actor successfully shifts the agenda*** (Nguyen 2015).

Figure 1: Visual representation of interactive communication structure and parameters of interest



Note: Figure displays text from the first 2016 general election presidential debate between Donald Trump and Hillary Clinton with moderator Lester Holt. Black font denotes observable corpus structure. Gray font denotes the additional latent structure in interactive communications, including an indicator for if turn t in debate d changed the topic ($l_{d,t}$) and each speaker m's ability to set the agenda (π_m). Gray dashed lines denote latent topic segment structure. Figure demonstrate Holt changing the topic ($l_{d,t} = 1$) and initiating a new topic segment on taxes.

2.3. *Defining features of an interactive political communication*

While political scientists frequently use text as data methods for topic discovery and measurement (see Grimmer and Stewart 2013), the discipline has yet to extend such methods to be suited for texts where multiple speakers *interact*. Such an extension is necessary because the text arising from a interaction is fundamentally different than the text usually studied in political science, namely because interactions are a social exercise in which speakers must negotiate who is speaking, when, and about what.

Figure 1 illustrates the defining features of text arising from an interaction using as an example the 2016 U.S. general election presidential debates. First, consider that the debates have a temporal structure. The words in the corpus can be grouped into uninterrupted utterances by a single speaker, or *speaking turns*, akin to each line in a transcript. Speaking turns contribute to the temporal structure of interactions as they have a natural, observable ordering. Further, since topics in an interactive communication ebb and flow, there's also latent temporal structure in the topics that arise. To see this, for each turn $t \in [1, T_d]$ in each debate $d \in [1, 3]$, let the binary latent variable $l_{d,t}$ indicate if the turn changed the topic ($l_{d,t} = 1$) or not ($l_{d,t} = 0$). All turns in which $l_{d,t} = 0$

that follow a topic-changing turn stay *on topic* and this sequence of speaking turns form a *topic segment*. Figure 1 demonstrates two topic segments. Initially the candidates discuss ISIS followed by a shift in topic to taxes. Gray dashed lines denote this latent topic segment structure.

In addition to the temporal structure that differentiates interactions from other political science corpora, interactions are also uniquely a social enterprise. First note that we observe multiple speakers in the corpus in Figure 1. Further, these speakers interact to influence how the text unfolds. To see this, Figure 1 denotes each speaker $m \in$ [Trump, Clinton, Holt] with a latent agenda setting ability, π_m. According to this model of speech, a speaker's agenda setting ability impacts how the text unfolds. Whether or not we observe a shift in topic when Clinton is speaking depends in part on how skilled she is at setting an advantageous agenda.

3 A MODEL OF AGENDA SETTING POWER

I measure the agenda setting power of actors in political interactions using the parametric Speaker Identity for Topic Segmentation (SITS) model (Nguyen, Boyd-Graber and Resnik 2012; Nguyen et al. 2014). Specifically, SITS builds upon Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan 2003) to account for and measure the temporal and social elements of interactions outlined in Figure 1, including where agenda shifts occur and the agenda setting power of actors (Nguyen 2015).

Next, I review the data generating process of a corpus under LDA and why this model poses limitations when applied to interactions. I then explain how SITS overcomes these limitations by extending LDA in order to measure agenda setting power of actors. Then with simulations I address two potential concerns about measuring agenda setting power with SITS. The simulations show that SITS performs well even with rapid dialogue typical of many interactions. The simulations also show that existing text as data approaches in political science do not perform well when tasked with measuring agenda setting in interactive settings. For ease of exposition, throughout this section, I will refer to any single interaction as a "discussion," each actor participating in a discussion as a "speaker," and each uninterrupted utterance by a speaker as a "speaking turn."

3.1. *From LDA to SITS*

To make the data generating process of LDA comparable to a text arising from an interaction, consider what is usually referred to as a "document" in LDA as single speaking turn in an interaction. Formally, for each discussion $d \in [1, D]$, consider each speaking turn $t \in [1, T_d]$ as a "document."[3] Given this alteration to the notation, the data generating process of LDA is as follows. First topics (ϕ_k), or probability distributions over the corpus vocabulary, are drawn for each of $k \in [1, K]$ topics from a symmetric Dirichlet distribution with parameter β. Then for each speaking turn $t \in [1, T_d]$ in each discussion $d \in [1, D]$, a distribution over topics ($\theta_{d,t}$) is drawn from a symmetric Dirichlet distribution with parameter α. Next, for each word index $n \in [1, N_{d,t}]$ in the speaking turn, a topic assignment ($z_{d,t,n}$) is drawn given the speaking turn's distribution over topics and a word ($w_{d,t,n}$) is drawn given its assigned topic.

LDA is not well-suited for discussion texts as it fails to capture the temporal and social dynamics of an interaction. LDA treats each document as a *new* mixture over topics; however, in a discussion, what the current speaker says is likely to be in response to the previous speaker's comments, thus the content of turn t is correlated with the content of speaking turn $t-1$. Further, LDA does not account for the social nature of discussions. Some speakers will exert more power over the latent topical agenda of a discussion than others.

SITS builds upon LDA to incorporate the temporal flow of discussion topics and agenda setting power of speakers into the data generating process outlined in Figure 2, where bold text indicates extensions to LDA. First, for each speaker $m \in [1, M]$, their agenda setting power (π_m) is drawn from a symmetric Beta distribution with parameter γ. As with LDA, K topics are drawn. Next, similar to LDA, a distribution over topics ($\theta_{d,t}$) must be assigned to each speaking turn $t \in [1, T_d]$ in discussion $d \in [1, D]$. However, this process unfolds differently for a discussion. SITS seeks to find sequences of speaking turns all on the same topic. Since the first turn of a discussion inherently changes the topic, this is noted by setting a turn-level topic shift binary variable equal

[3]Comparing data generating processes for texts from interactive and non-interactive settings is not straightforward. Alternatively, a "document" could be defined at the discussion-level, however this would disregard the interactive nature of discussions by obscuring all separate speaking turns into one instance of text.

Figure 2: Generative process of SITS

- **For each speaker $m \in [1, M]$, draw a speaker topic shift probability $\pi_m \sim \text{Beta}(\gamma)$.**
- For each topic $k \in [1, K]$, draw a topic-word distribution $\phi_k \sim \text{Dir}(\beta)$.
- For each turn $t \in [1, T_d]$, in each discussion $d \in [1, D]$ **(with speaker $a_{d,t}$)**:
  - **If $t = 1$, set the topic shift $l_{d,t} = 1$, otherwise draw $l_{d,t} \sim \text{Bernoulli}(\pi_{a_{d,t}})$.**
  - **If $l_{d,t} = 0$, set the topic distribution $\theta_{d,t} \equiv \theta_{d,t-1}$, otherwise** draw $\theta_{d,t} \sim \text{Dir}(\alpha)$.
  - For each word index $n \in [1, N_{d,t}]$:
    - Draw a topic $z_{d,t,n} \sim \text{Categorical}(\theta_{d,t})$.
    - Draw a word $w_{d,t,n} \sim \text{Categorical}(\phi_{z_{d,t,n}})$.

Note: Figure adapted Nguyen et al. (2014). The underlying data generating process of the parametric Speaker Identity for Topic Segmentation model. Bold text indicates extensions from Latent Dirichlet Allocation.

to one ($l_{d,t=1} = 1$). For all other turns, whether or not a topic change occurs is drawn from a Bernoulli distribution parameterized by the speaker's agenda setting measure ($\pi_{a_{d,t}}$, where $a_{d,t}$ is the observed speaker of the speaking turn). Therefore, whether or not a speaking turn changes the topic is influenced by its speaker's agenda setting power. If a topic change is indicated, a new topic distribution is drawn, otherwise the topic distribution from the previous turn carries over to the current turn ($\theta_{d,t} \equiv \theta_{d,t-1}$). Then, identical to LDA, topic assignments are drawn according to the speaking turn's distribution over topics and words are drawn according to topic assignments.

To estimate SITS models in all simulations and applications that follow, I use a Gibbs sampler written in Java by Viet An Nguyen that is available to the public (Nguyen 2014).[4]

3.2. *Simulations*

Two concerns might arise in regard to measuring the agenda setting power of actors in interactive settings using SITS. First, rapid dialogue and short speaking turns often characterize interactions, raising concerns over how well a topic modeling approach will perform. I address these concerns with a simulation study, and I find that SITS performs well even with rapid dialogue typical of many interactions. Second, one might think simpler automated text analysis methods could be leveraged to measure agenda setting. However, I show that standard text analysis methods used in political science do not perform well when adapted to the task of estimating the topic shifts

[4]Appendix A.1 provides additional details regarding the sampler. Appendix A.2 notes how I chose hyperparameters for the models.

that underly agenda setting behavior. Taken together, these simulations suggest SITS provides a valuable contribution to the suite of text as data methods political scientists have at their disposal.

3.2.1. *Parameter estimation with sparse texts*

Topic models utilize how words co-occur at the document level to discover latent topics, and therefore do not perform well with short documents (Hong and Davison 2010). In some interactive settings, discussion bounces rapidly from speaker to speaker, so each speaking turn (each "document") may feature few words. This seems to present a problem for using SITS to analyze the agenda setting power of speakers.

However, SITS estimates the latent topic segment structure within an interaction, stringing together sequences of speaking turns likely to be on the same topic. So rather than estimate a distribution over topics for *each speaking turn*, SITS estimates a distribution over topics for *each topic segment* allowing more data to inform topic discovery. Doing so should alleviate concerns about topic modeling with sparse texts. However, given this known limitation of topic models, it is pertinent to evaluate if SITS can measure agenda setting power accurately with short speaking turns.

To do so, I **simulated four discussion corpora** according to the data generating process outlined in Figure 2.[5] Each simulated corpus contained the same ten discussions between five speakers and fifteen topics. The only difference between the simulated corpora was the number of words per speaking turn, $N_{d,t}$, or how verbose the speakers were.[6] For each corpus, I increased the number of words in the speaking turns, using $N_{d,t} = [5, 10, 25, 50]$ words per speaking turn, respectively.[7]

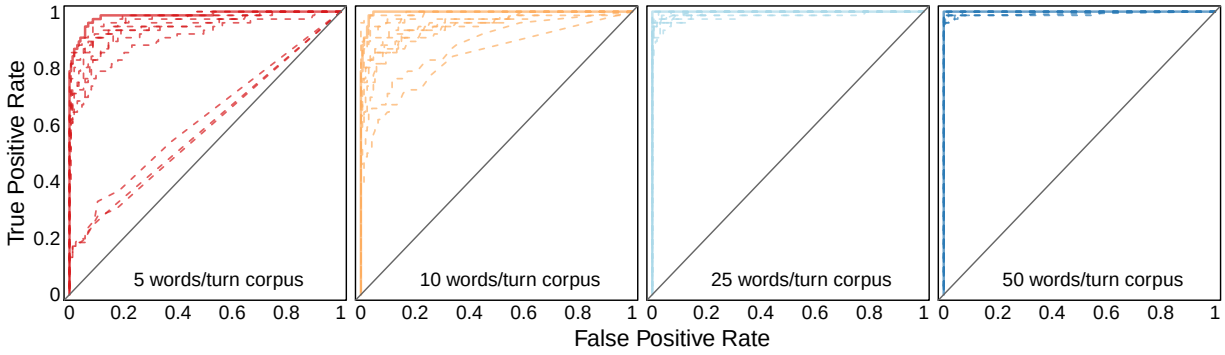
Then, to assess the ability of SITS to accurately estimate agenda setting power of speakers with sparse texts, I **estimated ten SITS models** for each simulated corpus. Rather than perfectly specify the model with known hyperparameter values, I estimated the ten models with *randomly*

[5]Simulating corpora according to the SITS data generating process is limited in assessing the applicability of the model to naturally occurring language that certainly deviates from this assumed process. To address this limitation, I present results in the Appendix C from a simulation exercise using observational data for which we do not know the true data generating process of the text. I find consistent results across simulations.

[6]As a consequence of varying speaking turn length, specific topic ($z_{d,t,n}$) and word ($w_{d,t,n}$) assignments varied, but the topic distributions ($\theta_{d,t}$) did not.

[7]Appendix B provides additional data-simulation details.

Figure 3: SITS recovers topic shift parameters with sparse conversation texts



Note: ROC curves for estimation of latent turn-level topic shift parameters. Plots show results for simulated corpora featuring 5, 10, 25, and 50 words per speaking turn, respectively. Each dashed line considers classification of one of ten estimated models per corpus. Bold line represents average classification across the ten models.

drawn hyperparameters.[8] This allows me to assess the performance of SITS under a more realistic scenario where researchers do not know the true value of hyperparameters, such as the number of topics K.

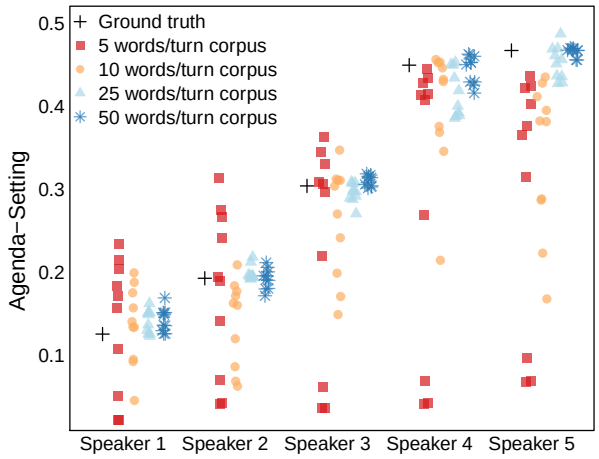
Figure 3 presents plots receiver operating characteristic (ROC) curves for the turn-level topic shift variables. ROC curves are a visualization of the diagnostic ability of a binary classifier—in this case, classifying turns as topics shifts or not—while varying the threshold at which to determine classification—in this case, varying the threshold at which a turn is considered a topic shift.[9] Each dashed line considers the classification ability of one of ten estimated models, and the bold lines represents classification after averaging across the ten estimated models. The x-axis is false positive rate and the y-axis is true positive rate. The model performs well at correctly identifying whether or not a topic change occurred, even with limited data of ten words per speaking turn, shown in the second plot. Even most of the models estimated using the five words per speaking turn corpus (in the first plot) perform well; however, with such limited data, the model is not able to overcome strong, unsuitable priors.[10] The model improves as it is provided more data, with almost perfect classification when the simulated data had 25 or 50 words per speaking turn.

[8] Appendix B provides model estimation details.

[9] I assess the posterior mean for the turn-level binary topic shift indicators.

[10] The three worst performing models were estimated with hyperparameter values of $\beta > .2$ which is far from the true value of .01. With more data, the model is able to overcoming such extreme hyperparameter values, as shown in the second, third, and fourth plots in Figure 3.

Figure 4: SITS recovers agenda setting parameters with sparse texts



Note: Estimates of speaker-level agenda setting parameters. Crosses represent the true value of the agenda setting parameter for each speaker. Squares, circles, triangles, and stars represent results from 10 estimated models for each speaker from simulated corpora featuring 5, 10, 25, and 50 words per speaking turn, respectively.

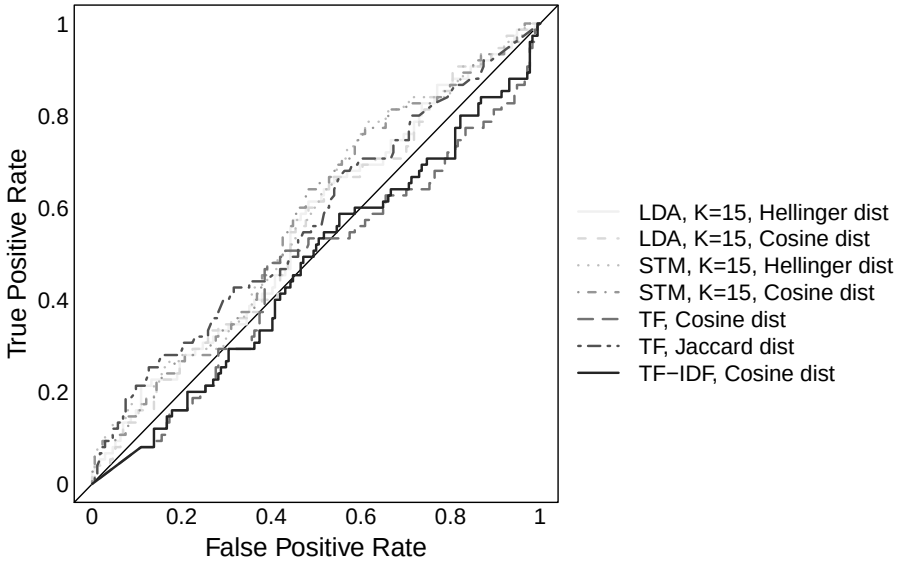
Figure 4 plots the true agenda setting measure for each speaker (crosses) and the estimated measures for each of the ten models as the number of words per speaking turn increases. Again, more data results in better estimation, as the agenda setting measurement is precise and accurate when the simulated data had 25 or 50 words per speaking turn. Figure 4 also provides reassurance that the model is robust to choice of hyperparameter, as accurate results are largely attained despite researcher-chosen hyperparameters.[11] However, the simulation suggests researchers should exercise caution with their hyperparameter selection when their texts feature few words per speaking turn. Figure 4 shows that when the simulated corpora had very few words per speaking turn and the models were estimated with poorly chosen hyperparameters, the model found few speaking turns that changed the topic and underestimated agenda setting power.

3.2.2. *Adapting standard methods*

Arguably, a simpler way to measure agenda setting abilities could be the following "two-step" process. First, detect topic shifts by measuring dissimilarity of consecutive speaking turns. That is, if turn t is dissimilar to turn $t-1$, it is likely the speaker of turn t shifted the topic. Second, calculate agenda setting power as, for example, the proportion of a speaker's turns that were classified as

[11] Appendix B provides details on hyperparameters used to estimate the models.

Figure 5: Standard text methods do not identify topic changes within interactions



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.

changing the topic. I next demonstrate that standard automated text analysis methods, when used in an interactive setting, do not reliably identify when speaking turns change the topic, and thus are not suited to measure the agenda setting power of actors.

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 consecutive speaking turns in several ways, including term frequency (TF) vectors, term frequency inverse document frequency (TF-IDF) weighted vectors, and with estimated latent topic proportions from the LDA and Structual Topic Models (STM) topic models (Roberts et al. 2014).[12] I then assess the classification of topic shifts with several distance metrics including cosine distance, Jaccard distance, and Hellinger distance.[13] I use the simulated 50 words per turn corpus, the largest simulated corpus, to provide these methods as much data as possible to detect shifts in topic.

Since these metrics do not have an intuitive threshold for determining if a topic shift occurred or

[12]The STM topic model included speaker and discussion indicators as prevalence covariates.

[13]Appendix B provides additional details on the simulation exercise.

not, I again use ROC plots to consider classification at any given threshold. Figure 5 plots the ROC curves. The two-step methods barely outperform random guessing as indicated by the diagonal black line. This simulation suggests that identifying speaking turns that change the topic—within the context of short texts typical of interactions—cannot be reliably accomplished with commonly used automated text analysis methods.[14]

4 APPLICATIONS

In the remainder of the article, I present three applications to demonstrate the validity and usefulness of the agenda setting measure of power. Taken together, the applications show that the measure captures the theoretical construct of agenda setting power, that the measure is valid across a diverse set of discursive settings, and that agenda setting correlates with important outcomes of political interactions.

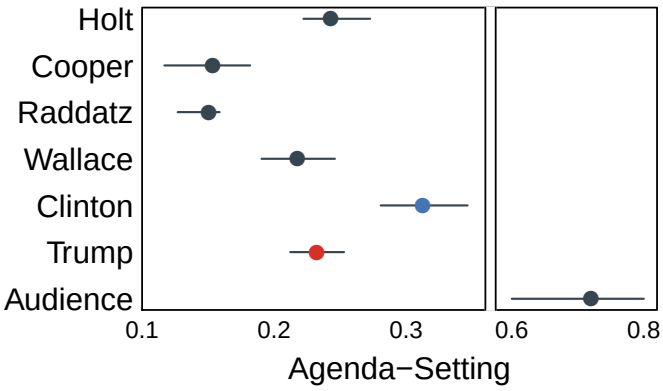
I first validate the measure with electoral debates, finding the agenda setting power of candidates aligns with media accounts of their performance. Then, I contrast agenda setting power with popular count-based measures of participation because political scientists often use such measures to quantify behavior in interactive settings. Using in-person deliberations, I find that agenda setting does not correlate with quantity of participation. I further show that agenda setting negatively correlates with attitude change—an important outcome of deliberations—while count-based measures fail to capture this pattern. Lastly, I conduct a more direct test of my central claim that the agenda setting is a form of power. Using a novel dataset of online deliberations where each participant is incentived to achieve their preferred outcome, I show that agenda setting power correlates with "winning" the deliberation.

4.1. *Validation exercise using electoral debates*

I first assess the validity of the agenda setting measure using the three 2016 U.S. presidential general election debates. I show that my measurement of the candidates' agenda setting power aligns with media accounts of their debate performances. Media accounts also validate my theoret-

[14] Appendix B provides simulation results for additional text methods. I find consistent results across all tested methods.

Figure 6: Agenda setting of debate participants



Note: Posterior mean agenda setting measures and 95% equal-tailed credible intervals for debate moderators, candidates, and audience members that strictly asked questions during the town-hall style debate.

ical argument that agenda setting power manifests through successfully shifting to and maintaining others' attention on advantageous topics. I additionally demonstrate how the latent topics estimated by the model can be used to explore the candidates' issue agendas.

The literature on how debates affect the mass public suggests that viewing presidential debates can increase issue knowledge and salience (see Benoit, Hansen and Verser 2003) and can help voters learn about candidates (e.g., Holbrook 1999). However, for a candidate to hope to garner these effects among viewers, candidates must make strategic choices during debates, specifically in regard to strategies they employ to set the debate's agenda (Boydstun, Glazier and Phillips 2013). SITS provides a way to measure and assess the agenda setting skills of electoral candidates.

In this application, I use transcripts from the 2016 U.S. presidential general election debates between Democratic nominee Hillary Clinton and Republican nominee Donald Trump that took place on September 26, October 9, and October 19, 2016 (Commission on Presidential Debates 2016). To measure the agenda setting power of the candidates and moderators, I estimated three SITS chains from the data with randomly drawn starting values.[15] I use iterations from all three chains to estimate posterior means of the agenda setting measures.

Figure 6 reports the estimated agenda setting power for the debate moderators, the candidates, and the audience members that strictly asked questions during the second town-hall style debate.

[15] Appendix **??** provides additional modeling details and convergence diagnostics.

Points are posterior means and bands are the 95% equal-tailed credible intervals. The model estimates Clinton was more successful at shifting the topic in a speaking turn than Trump, coming at little surprise as she has a reputation as a skilled debater (Chozick 2015), and Trump's campaign team struggled to convince him to practice for the debates (Healy, Parker and Haberman 2016).

Moreover, Figure 6 provides a source of validity for the agenda setting measure when coupled with media accounts of the candidates' performances. In regard to the first debate, panelists on a Fox News program, *Special Report with Bret Baier*, put the candidates' agenda setting measures into words (*Special Report with Bret Baier* 2016).[16] The panelist Bill McGurn of the Wall Street Journal said Clinton "did better than I expected" and that she put Trump "on defense a lot on his business stuff... He spent a lot of time defensive and explaining himself." McGurn describes the results in Figure 6—Clinton outperformed Trump at steering the debate toward her agenda. The next panelist, Caitlin Huey-Burns of RealClearPolitics, expressed a similar sentiment, saying "He missed a lot of opportunities to change the course of the debate back to what he's comfortable talking about... he didn't seem prepared to take these attacks and move on." Huey-Burns laments Trump's failure to set his agenda and tendency to detrimentally stayed *on* topic when disadvantageous topics were introduced. The third panelist, Monica Crowley of The Washington Times, arrived at a similar conclusion:

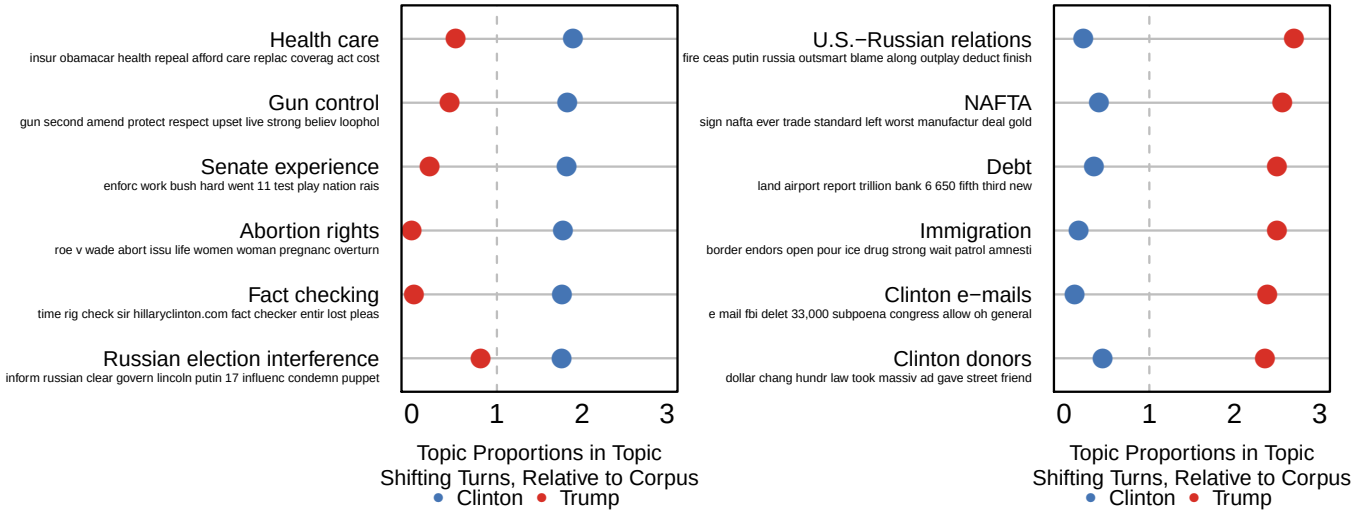
> "It's not helpful when he extends the life of a story that is not helpful to him... he should not have fallen for her bait. Clearly at the end of the debate she had that talking point prepared about women. And since Lester Holt didn't bring it up... she felt she needed to interject it... And it was a problem because he felt that then he had to address that."

Crowley notes Trump's inability to strategically set an advantageous agenda and specifically points to the memorable moment when Clinton shifted the topic to Trump's history of degrading women.[17] Overall, we see the sentiment of these Fox News contributors reflected in the candidates' agenda setting measures.

[16] I present opinions from panelists on a Fox News program, because as a conservative network, it should be the least critical of Trump. Appendix D.1 provides similar opinions of the candidates' performances across different media outlets.

[17] Appendix D.4 shows Clinton's comment is indeed a topic shift the model's results.

Figure 7: Candidate agendas in the 2016 presidential election debates



(a) Clinton shifted to health care, her political experience, and Russian election interference

(b) Trump shifted to foreign relations with Russia, Immigration, and Clinton's e-mails

Note: The x-axes show the topic proportions for topic segments shifted to by the candidate, relative to the topic proportions in the full corpus. The y-axes show top words selected using FREX weighting for the top six topics for each candidate.

Further, the moderators' low agenda setting measures may seem counterintuitive as a moderator's role is to pose new, topic-changing questions. However, the moderators' participation in the 2016 American presidential debates was namely in the form of enforcing time limits, allowing for responses, and re-asking questions when they are diverted—all participation that does not change the substantive topic of discussion, explaining their low agenda setting measures. An additional source of validity comes from the "Audience" participant in Figure 6. These participants posed questions to the candidates in the second town-hall style debate. Their only role was to change the topic, validating the high agenda setting measure.

Lastly, as a topic model, SITS allows the researcher to explore what topics candidates used their agenda setting power to promote. Issue ownership theory argues that voters associate certain issues with certain parties and suggests that electoral candidates will seek to discuss topics that they "own" and find advantageous (Petrocik 1996). SITS provides a means to discover debate topics and how candidates use agenda setting as a strategy to promote an advantageous agenda.

Figure 7 illustrates the candidates' agendas using the latent topics estimated from the model.

Specifically, the x-axes show the topic proportions for the topic segments the candidate was responsible for initiating, relative to the topic proportions in the full corpus.[18] The vertical line at 1 demonstrates when a topic is no more or less likely to be brought up by a candidate than it is brought up throughout the debates at large. The y-axes present top words for the six most shifted-to topics for each candidate. Top words were determined using FREX weighting, thus taking into account both the frequency and exclusivity of a word in a topic rather than the words with the highest probability of belonging to a topic (Bischof and Airoldi 2012).

Figure 7 shows that when setting the agenda during the debate, Clinton shifted to advantageous topics of health care reform and her political experience. She also shifted to disadvantageous topics for Trump—Russian election interference and calling on fact-checkers to assess Trump's claims during the debate. Compared to the extent to which these topics were discussed in the corpus at large, Clinton was twice as likely to discuss them when setting her agenda. On the other hand, Trump was more than twice as likely to discuss issues intended to harm Clinton—her use of a private e-mail server, her campaign donors, and the North American Free Trade Agreement (NAFTA) signed into law by Clinton's husband, former President Bill Clinton. While Clinton shifted to Russia's interference in the election, Trump shifted to the broader theme of foreign relations with Russia.

4.2. *Agenda setting and count-based participation measures*

Without a statistical model of discussion text equipped to measure latent speaker behaviors, researchers studying interactions across political contexts have resorted to measuring observable behaviors such as the number of words spoken by a participant. Therefore, I investigate how agenda setting power relates to status-quo measures of speaker behavior in an interactive setting. Importantly, I examine how count-based measures and agenda setting correlate with one important outcome of interactions—attitude change. I find that agenda setting negatively correlates with attitude change, while count-based measures do not, suggesting agenda setting is not identified

[18] For each candidate, I calculated topic proportions with topic assignments ($z_{d,t,n}$) from topic segments initiated by the candidate. I also calculated topic proportions for the entire corpus. The x-axes present the probability of discussing a topic when setting the agenda relative to the probability it is discussed across the entire corpus.

and accurately measured with count-based measures.

Deliberation texts were generously shared by Christopher Karpowitz and Hans Hassell from a pilot study examining the effect of stress on deliberation participation.[19] For this study, members of the Brigham Young University (BYU) community were recruited to discuss the BYU Dress and Grooming Standards, a specific set of rules governing the appearance of all students and staff at the university.

The study included ten discussion groups, each composed of four members. Participants first completed a pre-discussion survey about their attitudes regarding the Dress and Grooming Standards. Participants then engaged in a discussion in which they had 25 minutes to discuss the pros and cons of the standards and agree upon recommendations regarding changes to the standards, if any. Each group was instructed to write down up to two specific changes to the standards that they would then vote on after the discussion. Recommendations with a majority of the post-discussion votes would be sent to the Honor Code office, with no guarantee that the changes would be implemented. After the discussion, participants again reported their attitudes regarding the Dress and Grooming Standards. While not a political topic, the participants engaged in serious deliberations. They had to deliberate to share deeply held and sometimes conflicting perspectives, aggregate their individual preferences to two policy proposals, and vote on the proposals, making this a useful case study for the agenda setting measure.

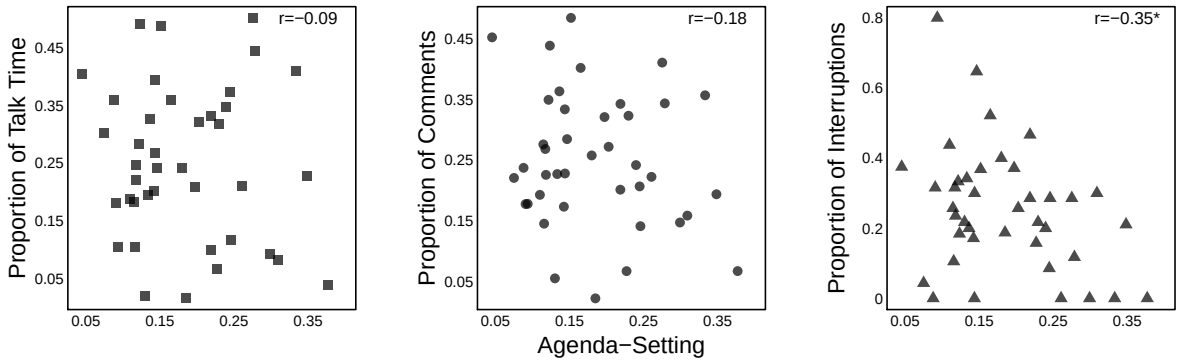
I estimated three SITS chains from the data with randomly drawn starting values.[20] I use iterations from all three chains to estimate posterior means of the agenda setting measures.

Figure 8 plots posterior mean estimates of agenda setting power against three commonly used count-based participation measures in the literature, including the proportion of speaking time used by a participant, the proportion of the group's comments made by a participant, and the proportion of the group's interruptions made by a participant. Each plot also displays the correlation coefficient, r, between the count-based measure and the agenda setting measure. First, there is no

[19] I find no evidence of a treatment effect; therefore, I do not include the treatment as a variable in subsequent analyses.

[20] Appendix E.2 provides additional modeling details and convergence diagnostics.

Figure 8: Agenda setting correlates with quality, not quantity, of participation



Note: *p<0.05. Figures report Pearson's r correlation coefficients. The y-axes are the speaker's proportion of group-level talking time, comments, and interruptions, respectively.

notable correlation between the first two count-based measures based on amount of participation. Second, we see a negative correlation between how often a participant interrupts others and their agenda setting tendency. Participants that interrupt less are also more likely to successfully advance their agendas, both signs of a higher-quality of participation. Taken together, these results suggest that the agenda setting measure captures something about the quality, rather than only the amount, of a speaker's participation.

Next, I examine how the count-based measures and the agenda setting measure correlate with attitude change. I assess the outcome of attitude change because an important question in the literature on political deliberation pertains to how deliberating affects one's attitudes, finding it can influence the formation and strength of issue attitudes (e.g., Levendusky, Druckman and McLain 2016; Klar 2014).

Before and after the discussion, participants were asked to rate their agreement with 23 questions regarding the purpose and fairness of the Dress and Grooming Standards on a seven point scale. All questions could be coded such that one indicated the least critical stance and seven indicated most critical stance toward the standards. To measure attitude change as my dependent variable, I calculated the absolute value of the mean difference between pre-discussion and post-discussion responses. Therefore, a value of .5 means the respondent changed their responses to *each* battery item, on average, by .5 points in one direction on the scale.

I then model the relationship between these different measures of speaker behavior during the

Table 1: Agenda setters less likely to change attitude on topic of deliberation

	Dependent variable: Attitude change			
	(1)	(2)	(3)	(4)
Proportion of talk time	−0.077			
	(0.188)			
Proportion of comments		0.014		
		(0.240)		
Proportion of interruptions			0.220	
			(0.199)	
Agenda setting				−0.635*
				(0.334)
Constant	0.258*	0.236*	0.184*	0.354*
	(0.053)	(0.061)	(0.056)	(0.080)
Observations	38	38	38	38
R^2	0.003	0.003	0.048	0.082

Note: *p<0.1. Coefficients from linear regressions with clustered standard errors at the discussion-group level in parentheses. Dependent variable is absolute value of a participant's attitude change regarding the Dress and Grooming Standards as indicated by the difference between pre- and post-discussion survey responses. Explanatory variables are those from Figure 8. Two observations omitted as outliers due to high Cook's distance.

deliberation and subsequent attitude change. Table 1 presents coefficients from four linear regression models with clustered standard errors at the group level in parentheses. I find no evidence of an effect of the count-based participation measures on attitude change, as the effect of each count-based measures is not distinct from zero in Models 1, 2, and 3. However, the negative coefficient on the agenda setting measure in Model 4 suggests that participants who succeeded at setting the agenda displayed less attitude change regarding the Dress and Grooming standards.

Perhaps agenda setters began the exercise with strongly held attitudes and strategically navigated the deliberation to effectuate their agendas. Or, it is possible that attitude strength played no role. Rather, as a simple consequence of setting the agenda, agenda setters heard less new information than others and had less of an opportunity to shift their views. Regardless, these results suggest agenda setting is an important behavior in the study of deliberations and that count-based participation metrics do not capture its impact on attitude change.

4.3. *Test of agenda setting as a form of power*

Lastly, I assess the validity of the agenda setting measure as a measure of power using a novel online deliberation study. The study incentivized participants to achieve their preferred outcome, or to to "win," the deliberation. I find that agenda setting correlates with winning the deliberation. Power over the deliberation's content correlates with power over the deliberation's outcome. Thus, this exercise directly tests a central argument of this article—that agenda setting is a form of power that is important to political interactions.

This study involved four stages. First, participants took a pretest survey. During the survey, participants learned[21] about five prominent charities and indicated which charity they wanted to receive a $1.00 donation from the researchers. The five charities were ALSAC - St. Jude Childrens Research Hospital, American Heart Association, American Red Cross, Doctors Without Borders USA, and UNICEF USA. Second, participants were randomized into partnerships conditional on disagreeing about which charity should receive the donation.[22] Third, participants engaged in a ten minute online, written deliberation with their assigned partner. Fourth, participants answered a short battery of post-deliberation questions.

This procedure yielded 61 online deliberations amongst Amazon Mechanical Turk participants collected between April and August 2019. Participants were paid $1.00 for the pretest survey and $3.00 for returning promptly to the deliberation. The participants were incentivized to "win" the deliberation—a participant received a $1.00 bonus if their charity was chosen to receive the researcher's donation in the post-deliberation survey by *both* participants. In addition to the bonus payment to the "winner," the participants were further incentivized to agree because the $1.00 would not be donated to any charity unless participants indicated agreement in the post-deliberation survey. Despite these incentives, three of the deliberations did not reach an agreement.

I estimated three SITS chains using all 61 deliberations with randomly drawn starting values.[23]

[21] Appendix F.1 shows charity information given to participants.

[22] Participants answered an open-ended question asking why they chose their preferred charity. Participants were not considered for the deliberation stage if this answer was of poor quality. Additionally, participants were not considered for the deliberation stage if, after learning that the task would be within the hour, they expressed they were not willing to return.

[23] Appendix F.4 provides additional modeling details and convergence diagnostics.

Table 2: Agenda setters more likely to achieve preferred deliberation outcome

	Dependent variable: Deliberation outcome	
	(1)	(2)
Agenda setting	3.79**	3.42*
	(1.77)	(1.66)
Constant	−.93	.98
	(1.52)	(1.16)
Partnership fixed effects	✓	
Controls		✓
Observations	116	116
AIC	274.02	177.5

Note: **p<0.5, *p<0.1. Coefficients from logistic regressions. Model 2 reports clustered standard errors at the partnership level in parentheses. Dependent variable is if participant won ($y = 1$) or lost ($y = 0$) the debate. Model 2 controls for the charity the participant argued for, gender, education, and ethnicity. Coefficient estimates for the control variables are reported in the Appendix.

I use iterations from all three chains to estimate posterior means of the agenda setting measures.

To assess if agenda setting during the deliberation correlates with "winning" (i.e., having one's preferred charity chosen by both participants post-deliberation), I estimate logistic regression models.[24] Table 2 presents results for the binary outcome of winning ($y = 1$) or losing ($y = 0$) the deliberation. The key explanatory variable is the agenda setting power of participant. Model 1 includes fixed effects for partnership. Model 2 includes control variables for the charity the participant argued for, gender, education, and ethnicity.[25] We see that the effect of agenda setting on who wins the deliberation is statistically significant in Model 1, and that the effect size is consistent across models. To put the coefficient estimate into context—moving from the first to the second Quartiles (or the second to the third) of agenda setting increases the probability of winning the deliberation by 40%.[26] This application provides evidence that agenda setting is a form of

[24] I first conducted a two-sided, paired Wilcoxon signed-rank test with the 58 deliberations that reached agreement. This is a non-parametric test useful for the paired (each deliberation has a winner and a loser) data. The average difference in means within partnership is .044 ($p = .125$), which is not strong enough evidence to conclude that agenda setting impacts deliberation outcomes. I turn to a logistic regression in order to model the outcome of the deliberation as a function of both agenda setting and additional covariates to reduce standard errors and increase the test's power.

[25] Appendix F.6 provides full model results.

[26] Quartiles of agenda setting are .080 (Zeroth quartile, minimum), .215 (First quartile), .299 (Second quartile), .406 (Third quartile), .833 (Fourth quartile, maximum).

power—participants who set the agenda get what they want out of the deliberation.

5 CONCLUSION

Power is a fundamental theoretical concept in the study of politics but remained difficult to quantify outside of the formal political arena where votes, vetoes, and decision-making in general are observed. In this article, by considering the role of power in one of the most basic political actions—talking with others—I've introduced a measure of power applicable to the countless interactions that occur across the political sphere.

Importantly, I validated the agenda setting measure across a diverse set of discursive setting. Debates are oppositional, strategic, and have the goal of identifying a "winner" and "loser." Conversely, deliberations are characterized by collaboration and thoughtful consideration of all perspectives. I validated the agenda setting measure of power in both settings. Further, I validated the measurement with both in-person and online settings. Interactions unfold differently online without body language or other interpersonal cues to guide the communication. Moreover, transcripts from online and in-person interactions differ as online communications feature slang, symbols, and typos. Validating the agenda setting measure with both in-person and online interactions demonstrates its utility across these different settings.

There are of course limitations to the measure of agenda setting power proposed here. First, this measure does not specify if it was *what* the person said or *how* they said it that changed the course of the agenda. Perhaps someone's serious or humorous tone shifted the room's attention to their point, rather than how interesting or persuasive it was—this measure can not distinguish how the change in topic was achieved. Relatedly, this measure of power is not able to capture deference to unspoken, but understood opinions of the most powerful person in the room. If a CEO attends a meeting between their employees, she is likely exerting power over what her employees say, even if the she remains silent. The agenda setting measure overlooks the latent power of the CEO and fails to capture of the overarching power dynamics of this scenario.

Moreover, SITS could be extended to measure additional latent structures within political interactions. For example, the SITS data generating process could be altered to account for a speaker's

tendency to bring up similar topics over time. One might suspect Clinton is likely to bring up Russian interference in the election because she shifted to this topic earlier in the debate. If SITS is extended to account for the same person shifting to similar topic distributions, then one could further imagine detecting a latent *coalition* of speakers that shift to similar topic distributions.

The results presented in this article are intended to validate the agenda setting measure and stress its importance to the study of political interactions. Teasing out the theoretical role of agenda setting in interactions is left for future work. One open question is, under what conditions does agenda setting power equate to perceived power? Perhaps others tend to *discount* a woman's power over the communication when she sets the agenda. Or, depending on the social norms of the setting, setting the agenda could be perceived as rude, controlling, or irritating to others, so exercising agenda setting power may harm an individual's perceived power. Understanding the relationships between agenda setting power, perceived power, influence, persuasion, and more is left for future work.

Finally, SITS is just one of many automated topic segmentation methods. SITS is particularly useful in an interactive setting because it uses speaker identity to inform the segmentation task. However, other topic segmentation methods have promise for researchers needing to find coherent topic segments beyond interactions (see Purver 2011). For example, it might be useful to separate a Twitter user's timeline—a sequence of *short* texts—into topic segments to see when and why the user discusses particular topics. Further, topic segmentation might be useful for researchers trying to segment a *large* document into more manageable sizes or theoretically useful quantities. This might be needed for manual coding by crowd sourced workers, for example. Topic segmentation methods provide a principled way to approach tasks like these where the text is not already structured in the most theoretically or practically useful way.

This article set out to quantify the power dynamics that lie under the surface anytime two people talk. Power in debates, discussions, and deliberations is often overlooked by empirical researchers because we lack of methodological tools suited for these settings. This article helps shift the quantitative study of political power to interactions by offering a technique for measuring

the important concept of power across a variety of interactive settings. The hope is that the measure provides opportunity for additional theoretical development and principled analysis regarding the power dynamics in political interactions.

5 References

Bachrach, Peter and Morton S Baratz. 1962. "Two Faces of Power." *American Political Science Review* 56:947–952.

Barabas, Jason. 2004. "How Deliberation Affects Policy Opinions." *American Political Science Review* 98(4):687–701.

Beck, Paul Allen, Russell J Dalton, Steven Greene and Robert Huckfeldt. 2002. "The Social Calculus of Voting: Interpersonal, Media, and Organizational Influences on Presidential Choices." *American Political Science Review* 96(1):57–73.

Benoit, William L, Glenn J Hansen and Rebecca M Verser. 2003. "A Meta-Analysis of the Effects of Viewing US Presidential Debates." *Communication Monographs* 70(4):335–350.

Bischof, Jonathan and Edoardo M Airoldi. 2012. Summarizing Topical Content with Word Frequency and Exclusivity. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. pp. 201–208.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(Jan):993–1022.

Boydstun, Amber E, Rebecca A Glazier and Claire Phillips. 2013. "Agenda Control in the 2008 Presidential Debates." *American Politics Research* 41(5):863–899.

Boydstun, Amber E, Rebecca A Glazier and Matthew T Pietryka. 2013. "Playing to the Crowd: Agenda Control in Presidential Debates." *Political Communication* 30(2):254–277.

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.

Chozick, Amy. 2015. "In Debate, Hillary Clinton Will Display Skills Honed Over a Lifetime." *New York Times*. October 9.
https://www.nytimes.com/2015/10/11/us/politics/in-debate-hillary-clinton-will-display-skills-honed-over-a-lifetime.html. Last accessed on September 20, 2019.

Chozick, Amy and Michael Barbaro. 2016. "Hillary Clinton, Mocking and Taunting in Debate, Turns the Tormentor." *New York Times*. October 20.
https://www.nytimes.com/2016/10/20/us/politics/hillary-clinton-donald-trump.html. Last accessed on September 25, 2019.

Commission on Presidential Debates. 2016. "Debate Transcripts." *debates.org*.
https://www.debates.org/voter-education/debate-transcripts/. Last accessed on September 20, 2019.

Dahl, Robert A. 1957. "The Concept of Power." *Behavioral Science* 2(3):201–215.

Douthat, Ross. 2016. "Trump Fails the Stamina Test." *New York Times*. September 27.

https://www.nytimes.com/interactive/projects/cp/opinion/clinton-trump-first-debate-election-2016/trump-fails-the-stamina-test. Last accessed on September 25, 2019.

Druckman, James N and Kjersten R Nelson. 2003. "Framing and Deliberation: How Citizens' Conversations Limit Elite Influence." *American Journal of Political Science* 47(4):729–745.

Edsall, Thomas B. 2016. "No More Mr. Nice Guy." *New York Times*. September 27. https://www.nytimes.com/interactive/projects/cp/opinion/clinton-trump-first-debate-election-2016/no-more-mr-nice-guy. Last accessed on September 25, 2019.

Eggers, Andrew C and Arthur Spirling. 2016. "The Shadow Cabinet in Westminster Systems: Modeling Opposition Agenda Setting in the House of Commons, 1832–1915." *British Journal of Political Science* pp. 1–25.

Epstein, Lee, William M Landes and Richard A Posner. 2010. "Inferring the Winning Party in the Supreme Court from the Pattern of Questioning at Oral Argument." *The Journal of Legal Studies* 39(2):433–467.

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.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267–297.

Healy, Patrick, Ashley Parker and Maggie Haberman. 2016. "New Debate Strategy for Donald Trump: Practice, Practice, Practice." *New York Times*. September 28. https://www.nytimes.com/2016/09/29/us/politics/donald-trump-debate.html. Last accessed on September 20, 2019.

Holbrook, Thomas M. 1999. "Political Learning from Presidential Debates." *Political Behavior* 21(1):67–89.

Hong, Liangjie and Brian D Davison. 2010. Empirical Study of Topic Modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*. ACM pp. 80–88.

Huckfeldt, Robert, Paul E Johnson and John Sprague. 2004. *Political Disagreement: The Survival of Diverse Opinions within Communication Networks*. Cambridge University Press.

Johnson, Timothy R, Paul J Wahlbeck and James F Spriggs. 2006. "The Influence of Oral Arguments on the US Supreme Court." *American Political Science Review* 100(1):99–113.

Karpowitz, Christopher F, Tali Mendelberg and Lee Shaker. 2012. "Gender Inequality in Deliberative Participation." *American Political Science Review* 106(3):533–547.

Kathlene, Lyn. 1994. "Power and Influence in State Legislative Policymaking: The Interaction of Gender and Position in Committee Hearing Debates." *American Political Science Review* 88(3):560–576.

Klar, Samara. 2014. "Partisanship in a Social Setting." *American Journal of Political Science* 58(3):687–704.

Levendusky, Matthew S, James N Druckman and Audrey McLain. 2016. "How Group Discussions Create Strong Attitudes and Strong Partisans." *Research & Politics* 3(2):1–6.

Lukes, Steven. 1974. "Power: A Radical View." *London, New York* .

Mason, Melanie. 2016. "The Most Memorable Moments from the First Presidential Debate." *Los Angeles Times*. September 26.
https://www.latimes.com/politics/la-na-pol-debate-moments-20160926-snap-htmlstory.html.
Last accessed on September 25, 2019.

Mikhaylov, Slava, Michael Laver and Kenneth R Benoit. 2012. "Coder Reliability and Misclassification in the Human Coding of Party Manifestos." *Political Analysis* 20(1):78–91.

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* .

Mutz, Diana C. 2006. *Hearing the Other Side: Deliberative versus Participatory Democracy*. Cambridge University Press.

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. 2015. "Guided Probabilistic Topic Models for Agenda-Setting and Framing." PhD dissertation. University of Maryland, College Park.
https://drum.lib.umd.edu/handle/1903/16600.

Nguyen, Viet-An, Jordan Boyd-Graber and Philip Resnik. 2012. SITS: A Hierarchical Nonparametric Model Using Speaker Identity for Topic Segmentation in Multiparty Conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics pp. 78–87.

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.

Petrocik, John R. 1996. "Issue Ownership in Presidential Elections, with a 1980 Case Study." *American Journal of Political Science* pp. 825–850.

Prior, Markus. 2009. "The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure." *Public Opinion Quarterly* 73(1):130–143.

Purver, Matthew. 2011. "Topic Segmentation." *Spoken Language Understanding: Systems for Extracting Semantic Information From Speech* pp. 291–317.

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

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.

Ross, Janell. 2016. "Trump on 'fat slobs,' housekeepers and women who dont have that 'presidential look'." *The Washington Post*. September 27. https://www.washingtonpost.com/news/the-fix/wp/2016/09/27/trump-on-fat-slobs-housekeepers-and-women-who-dont-have-that-presidential-look/. Last accessed on September 25, 2019.

Schattschneider, Elmer E. 1975. *The Semi-Sovereign People: A Realist's View of Democracy in America*. Wadsworth Publishing.

Stanford Network Analysis Project. 2013. "Amazon Fine Foods reviews." http://snap.stanford.edu/data/web-FineFoods.html. Last accessed September 20, 2019.

Special Report with Bret Baier. 2016. "First presidential debate takeaways." *Fox News*. September 27. https://www.foxnews.com/transcript/first-presidential-debate-takeaways.

Wehner, Peter. 2016. "Birth of Trump News Network." *New York Times*. October 9. https://www.nytimes.com/interactive/projects/cp/opinion/clinton-trump-second-debate-election-2016/birth-of-the-trump-news-network. Last accessed on September 25, 2019.

Appendix: *Measuring Agenda Setting Power in Interactive Political Communications*

A ESTIMATION

A.1 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 ($l_{d,t}$).

A.2 A note on hyperparameters

Each corpus used in this paper—presidential debates, lab deliberations, and online deliberations—is different in length, number of speakers, and variation in how many different topics are covered within and across the set of interactions. I chose hyperparameters for modeling each corpus based on my prior expectation of how the topics (hyperparameter α) and distribution over topics (hyperparameter β) should look. In all cases, I chose α and β to induce sparsity in the topic-word and document-topic distributions, respectively. I chose $\gamma = 1$ in all cases to have a uniform prior over speaker agenda setting parameters.

Lastly, I chose K based on my understanding of the topical content covered in each corpora after several readings of each corpora. The lab deliberations and online deliberations each featured the same small number of topics across each individual deliberation, whereas the debates covered many topics, some reappearing across all three debates, but others not. I tried a few different choices of K, increasing it until I was satisfied that each topics (via high probability and FREX top words) was capturing a single (non-overlapping) concept. and that each topic captured a recognizable topic from the content of the corpus.

Specific hyperparameter choices are outlined in Appendix D.2 (presidential debates), Appendix E.2 (in-person deliberations), and Appendix F.4 (online deliberations).

B SIMULATION STUDY DETAILS

Each discussion had 25 speaking turns, and each turn was randomly assigned a speaker. True speaker agenda setting measures were set at intervals equally spaced by .1 from .1 to .5, rather than

drawn randomly, to assess the model's ability at recovering agenda setting measure at different points on the scale. Each topic was drawn from a Dirichlet distribution with symmetric $\beta = .01$ over a vocabulary of length 1000. As per the data-generating process in Figure 2, whether or not a speaking turn changed the topic was determined by the speaker's agenda setting measure. If a topic change was indicated, a new topic distribution over the speaking turn was drawn from a Dirichlet distribution with symmetric $\alpha = .1$. Topic assignments for each word in the speaking turn were drawn from the turn's topic distribution, and word indices were drawn given the topic assignments.

For each simulated corpus, I estimated ten models with randomly drawn hyperparameters. The hyperparameters used to estimate the models were drawn according to $\alpha \sim \text{Uniform}(0, .3)$, $\beta \sim \text{Uniform}(0, .3)$, and $\gamma \sim \text{Uniform}(0, 2)$ to exaggerate the range of values researchers usually pick from for these hyperparameters in topic modeling. Additionally, the number of topics was randomly drawn from a wide range $K \in [5, 25]$. The specific parameterizations were the following:

	α	β	γ	K
Model 1	.08	.049	1.248	6
Model 2	.23	.256	1.656	20
Model 3	.097	.055	.976	6
Model 4	.101	.248	.529	9
Model 5	.104	.205	1.278	5
Model 6	.147	.278	.086	17
Model 7	.159	.076	.316	13
Model 8	.212	.042	1.812	19
Model 9	.223	.043	.51	5
Model 10	.261	.267	.057	11

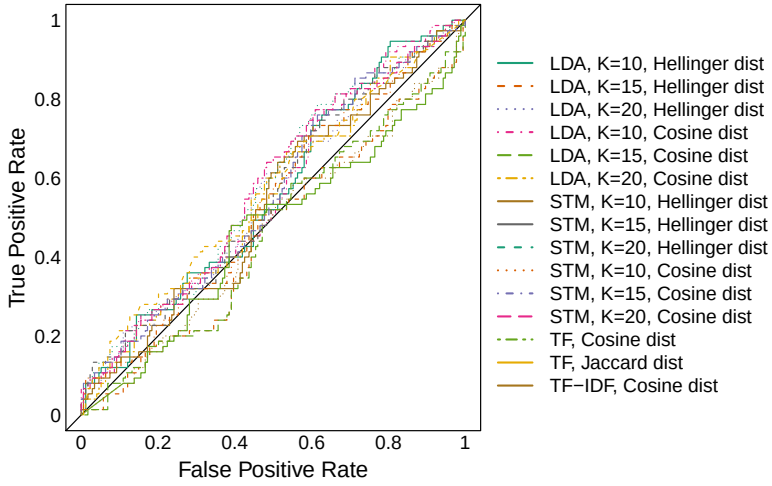
Each model ran for 100,000 iterations with 95,000 burn-in iterations.

Figure 1 presents similar results as in the main body of the article with additional model specifications (varying choice of K) and additional distance metrics. The additional simulation tests similarly show standard text as data methods in the field are not adaptable to the task of identifying topic shifts in short texts.

C SIMULATION STUDY WITH REAL TEXTS

As an additional simulation study, I simulate a corpus with real-world text for which I do not know the true data generating process of the *words and topics*. Importantly, I first generate the *structure* of the corpus according to the SITS data generating process. For five speakers, the agenda setting measures were set at intervals equally spaced by .1 from .1 to .5, rather than drawn randomly, to assess the model's ability at recovering agenda setting measure at different points on the scale. Then for 50 speaking turns in 5 discussions, I randomly draw a speaker for each turn,

Figure 1: Standard text methods fail to identify topic changes in simulated conversation



Note: Figure presents ROC curves for turn-level topic shift classification. Adapted text as data methods do hardly better or worse than random guessing (indicated by diagonal gray line), regardless of chosen threshold.

and conditional on the speaker's agenda setting power, I draw if a topic change occurred for that speaking turn.

Then, given this structure, I fill in the corpus with text. I use a collection of Amazon product reviews in order to do so (Stanford Network Analysis Project 2013). Specifically, I use a collection of reviews on four products—an oatmeal cookie, barbecue chips, a dog treat dispensing toy, and gluten free pancake mix. Moving sequentially through the corpus, when a topic change is indicated, I randomly draw which "topic" the topic segment should discuss (i.e., which Amazon product). Then, for each speaking turn in that topic segment (i.e., for subsequent turns that do not change the topic), I randomly a draw review for that product. The idea is to let a single review represent a single speaking turn. The resulting corpus mimics the ebb and flow of topics in a discussion.

This simulation procedure ensures I know the true values of the parameters of interest (turn-level topic change indicators and speaker-level agenda setting parameters). However, I am able to avoid assuming the data generating process for the topics (ϕ_k), turn-level distributions over topics ($\theta_{d,t}$), word-level topic assignments ($z_{d,t,n}$), and the words ($w_{d,t,n}$) as I did in the simulation in the main body of the article.

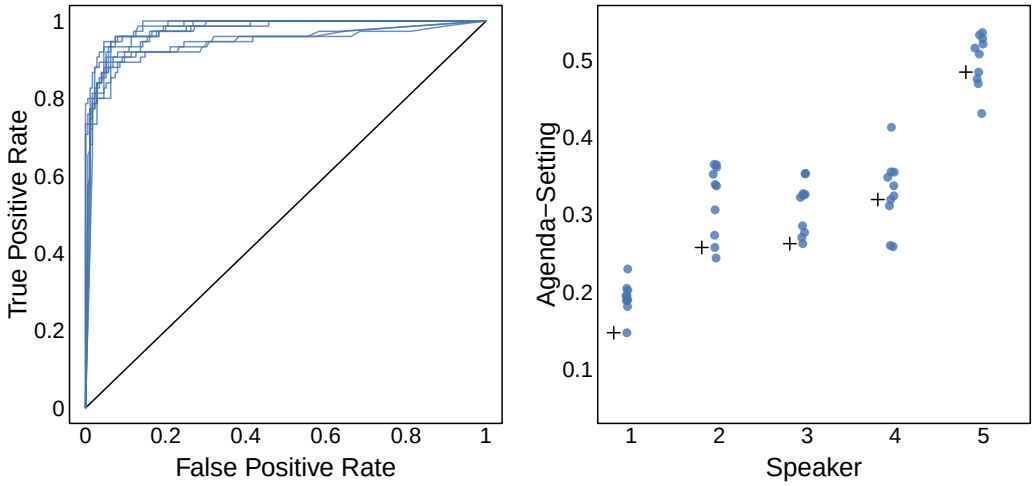
In order to make the product reviews seem like speaking turns in a discussion, I first subset to reviews with less than 500 characters (this leads to reviews with 100-150 words). Then, I preprocess the text by transforming it to lowercase, removing stopwords, removing numbers, removing punctuation, stemming, and removing any HTML. Finally, I removed any terms that were in only one review.

With corpus in hand, I then estimate ten SITS models with randomly drawn starting values

and hyperparameters. Because I do not know the true hyperparameter values for this simulation, I attempt to draw random values from reasonable ranges.

- $\alpha \sim \text{Unif}(0, .2)$
- $\beta \sim \text{Unif}(0, .05)$
- $\gamma \sim \text{Unif}(0, 2)$
- $K \in [6, 15]$
- 100,000 total iterations with 95,000 burn-in

Figure 2: Simulation exercise with real text



Note: First plot shows ROC curves for estimation of latent turn-level topic shift parameters. Each line considers classification of one of ten estimated models. Second plot shows estimates of speaker-level agenda setting parameters. Crosses are true values, and each point considers estimate from of one of ten estimated models. Models are estimated with varying hyperparameter values for the same simulated corpus.

The results for the SITS estimations are presented in Figure 2. Seven models achieve a high true positive rate and low false positive rate, while three of the models perform slightly worse. The resulting agenda setting posterior means are similarly accurate.

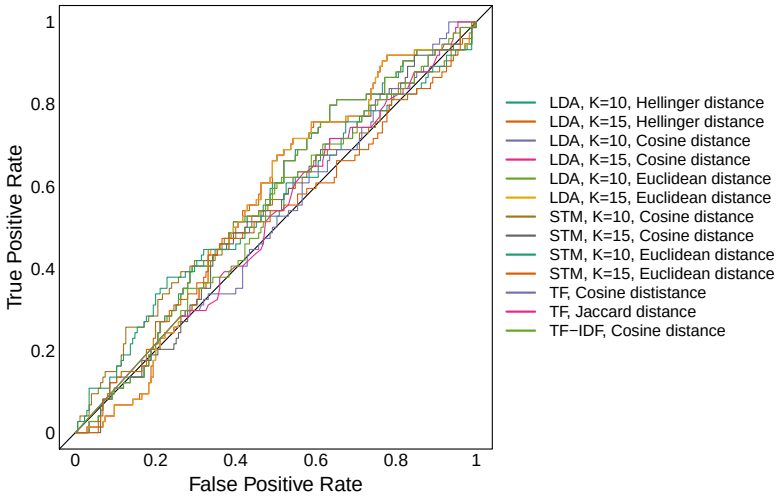
As in the main body of the article, I next attempt to find topic changes with several commonly used text methods. Figure 3 presents results for these methods. Again, we see that the methods do no better than randomly guessing which speaking turns change the topic.

D PRESIDENTIAL DEBATES

D.1 Media reaction to candidates' performance

Commentary on the first debate in the *New York Times* opinion section echoed that of the Fox News commentators. Columnist Ross Douthat said "[Trump] showed no ability to evade or

Figure 3: Standard text methods fail to identify topic changes in simulated conversation



Note:Figure presents ROC curves for turn-level topic shift classification. Adapted text as data methods do hardly better or worse than random guessing (indicated by diagonal gray line), regardless of chosen threshold.

duck or simply retreat on issues—his business dealings and his taxes, birtherism and racism—where long Trumpish answers make things only worse. Instead he kept over-litigating things and trying clumsy forms of jujitsu (Hillary's the real birther, etc.) that dragged bad moments out" (Douthat 2016). Similarly, Thomas Edsall wrote about the final moments of the first debate saying "The theme of Trump's struggle to express himself verses Clinton's forceful presentation built to a crescendo throughout the 90 minutes, culminating in the final minutes when Clinton became the clear aggressor" (Edsall 2016).

Commentary on the second debate did applaud Trump's preparation, while still noting Clinton's superior performance. Columnist Peter Wehner said "Trump did scores some points — he seemed somewhat more focused than in the past — and Clinton wasn't as consistently sharp and effective as she was in the first debate. But she certainly did better than he did, and Clinton won the key moments" (Wehner 2016).

Commentary on the third debate mentioned that Trump again "repeatedly gave up chances to respond to pointed taunts from Mrs. Clinton, who dominated the confrontation from its opening moments, needling and baiting him over and over" (Chozick and Barbaro 2016). Further, Chozick and Barbaro mention that "During a back-and-forth about immigration, Mrs. Clinton landed a hard jab, asserting that Mr. Trump had used undocumented workers to build Trump Tower even threatening such workers, she said, with deportation. Mr. Trump... did not respond" and shortly later, they mention, "...he allowed Mrs. Clinton to avoid entirely a question from the moderator" (Chozick and Barbaro 2016).

D.2 Text preprocessing

To preprocess the raw debate transcripts I removed punctuation, transformed all words to lowercase, removed stop words, and stemmed. I also removed all tokens that did not occur in three or more speaking turns. Finally, I removed 17 tokens that occurred more than 125 times in the corpus. The corpus vocabulary then contained 1,577 tokens across 933 speaking turns.

D.3 Convergence and topic alignment

I estimated three models with hyperparameters $K = 70$, $\alpha = 0.025$, $\beta = .01$, $\gamma = 1$ and different starting values. 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 of 1 with probability .2 for the first model, .1 for the second model, and .067 for the third model. The chains ran for 3,00,000. I discard the first 2,000,000 as burn-in iterations. Moreover, because of the high-dimensional nature of the model with 70 topics, I collect only every 100^{th} iteration of the $z_{d,t,n}$ topic assignment variables.

The correlations of the posterior means for the topic shift indicators across the thee models are (all $p < .001$):

	Model 1	Model 2	Model 3
Model 1	1	.90	.85
Model 2		1	.81
Model 3			1

The correlations of the posterior means for the agenda setting measures across the thee models are (all $p < .001$):

	Model 1	Model 2	Model 3
Model 1	1	1	.99
Model 2		1	1
Model 3			1

The correlations are simple assessment (to avoid plotting chains of over 900 parameters) of between-chain agreement. We see that the posterior means from models one and two are most highly correlated. However, I also conduct formal convergence diagnostics.

The Geweke diagnostic tests for equality of the means of the first and last part of a chain. I test the first 50% and last 50% of the turn-level binary topic shift indicators indicated 86%, 89%, and 88% of the 933 topic-shift variable chains passed the test, respectively. Further, using the Gelman diagnostic (Gelman and Rubin 1992; Brooks and Gelman 1998), I find that $R_c < 1.2$ is reached for 90% of the 933 topic shift indicators, thus I am fairly confident convergence was reached.

Lastly, I do not formally assess convergence of the topic assignment latent variables estimated by the model $(z_{d,t,n})$ for two reasons discussed in Section E.2. Rather, I assess topic coherence and alignment. Topic coherence, when considering the $M = 10$ most probable words for each topic, is

consistent across models. The total coherence for all 70 topics across the 3 models is -5212.071, -5196.303, and -5133.359, respectively. Model three best optimizes this coherence metric.

Next, I assess topic alignment, described in Section E.2. We see that the median number of shared words between the aligned topics of models one and two, when considering all 70 aligned topics, is 7 words. However, the median number of shared words between model three and the others is 5. Because of these results coupled with close readings of the transcripts and assessment of FREX top words, I report model one's FREX top words in Figure 7 in the main body of the article.

Candidate	Reference	Average $L1$	Median number of shared top words
1	2	0.68	7
1	3	0.80	5
2	1	0.67	7
2	3	0.83	5
3	1	0.79	5
3	2	0.85	5

D.4 Example of topic changes

Table 3 presents an example of the estimation of topic shifts. The first column shows the posterior mean of the topic shift indicator variable. Note that I do not consider turns with 4 or less tokens able to change the topic, thus $l_{d,t}$ is not sampled and is set to 0. The second column displays a sequence of turns before preprocessing. This 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 thus we see several topic changes occur, including the instance in which Clinton shifts the topic to Trump's history of insulting women.

Table 3: Example of topic changes in 2016 presidential debate

P(Shift)	Speaking Turn
0.73	***Holt*** Mr. Trump, this year Secretary Clinton became the first woman nominated for president by a major party. Earlier this month, you said she doesn't have, quote, "a presidential look." She's standing here right now. What did you mean by that?
0.01	***Trump*** She doesn't have the look. She doesn't have the stamina. I said she doesn't have the stamina. And I don't believe she does have the stamina. To be president of this country, you need tremendous stamina.
0.00	***Holt*** The quote was, "I just don't think she has the presidential look."
1.00	***Trump*** You have—wait a minute. Wait a minute, Lester. You asked me a question. Did you ask me a question? You have to be able to negotiate our trade deals. You have to be able to negotiate, that's right, with Japan, with Saudi Arabia. I mean, can you imagine, we're defending Saudi Arabia? And with all of the money they have, we're defending them, and they're not paying? All you have to do is speak to them. Wait. You have so many different things you have to be able to do, and I don't believe that Hillary has the stamina.
0.00	***Holt*** Let's let her respond.
0.76	***Clinton*** Well, as soon as he travels to 112 countries and negotiates a peace deal, a cease-fire, a release of dissidents, an opening of new opportunities in nations around the world, or even spends 11 hours testifying in front of a congressional committee, he can talk to me about stamina.
1.0	***Trump*** The world—let me tell you. Let me tell you. Hillary has experience, but it's bad experience. We have made so many bad deals during the last—so she's got experience, that I agree. But it's bad, bad experience. Whether it's the Iran deal that you're so in love with, where we gave them $150 billion back, whether it's the Iran deal, whether it's anything you can—name—you almost can't name a good deal. I agree. She's got experience, but it's bad experience. And this country can't afford to have another four years of that kind of experience.
0.00	***Holt*** We are at—we are at the final question.
0.00	***Clinton*** Well, one thing. One thing, Lester.
0.00	***Holt*** Very quickly, because we're at the final question now.
1.00	***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...
0.00	***Trump*** I never said that.
0.00	***Clinton*** .... women don't deserve equal pay unless they do as good a job as men.
0.00	***Trump*** I didn't say that.
0.03	***Clinton*** And one of the worst things he said was about a woman in a beauty contest. He loves beauty contests, supporting them and hanging around them. And he called this woman "Miss Piggy." Then he called her "Miss Housekeeping," because she was Latina. Donald, she has a name.
0.00	***Trump*** Where did you find this? Where did you find this?
0.00	***Clinton*** Her name is Alicia Machado.
0.00	***Trump*** Where did you find this?
0.00	***Clinton*** And she has become a U.S. citizen, and you can bet...
0.00	***Trump*** Oh, really?
0.00	***Clinton*** ... she's going to vote this November.
0.00	***Trump*** OK, good. Let me just tell you...
0.41	***Holt*** Mr. Trump, could we just take 10 seconds and then we ask the final question...
0.49	***Trump*** You know, Hillary is hitting me with tremendous commercials. Some of it's said in entertainment. Some of it's saidsomebody who's been very vicious to me, Rosie O'Donnell, I said very tough things to her, and I think everybody would agree that she deserves it and nobody feels sorry for her. But you want to know the truth? I was going to say something...
0.00	***Holt*** Please very quickly.
0.99	***Trump*** ...extremely rough to Hillary, to her family, and I said to myself, "I can't do it. I just can't do it. It's inappropriate. It's not nice." But she spent hundreds of millions of dollars on negative ads on me, many of which are absolutely untrue. They're untrue. And they're misrepresentations. And I will tell you this, Lester: It's not nice. And I don't deserve that. But it's certainly not a nice thing that she's done. It's hundreds of millions of ads. And the only gratifying thing is, I saw the polls come in today, and with all of that money...
0.00	***Holt*** We have to move on to the final question.

Note: A sequence of turns from the first 2016 general election debate. The first column is the posterior means of topic shift indicator after the burn-in period, the second column is raw debate text, and dotted lines indicate topic segments.

E IN-PERSON DELIBERATIONS

E.1 Text preprocessing

There were ten deliberations regarding the BYU Dress and Grooming Standards. On average, there are 90 speaking turns per deliberation. To preprocess the text I removed punctuation, transformed all words to lowercase and stemmed. Finally, I removed all tokens that did not occur in two or more speaking turns. The corpus vocabulary then contained 1,902 words across 899 speaking turns.

E.2 Convergence and topic alignment

I estimated three models with hyperparameters $K = 18$, $\alpha = .1$, $\beta = .1$, $\gamma = 1$ and different starting values. 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 of 1 with probability .05 for the first model, .1 for the second model, and .25 for the third model. The chains ran for 1,500,000. I discard the first 500,000 as burn-in iterations. I do not estimate $l_{d,t}$ parameters when the speaking turn has less than 5 tokens.

The correlations of the posterior means for the topic shift indicators across the thee models are (all $p < .001$):

	Model 1	Model 2	Model 3
Model 1	1	.97	.96
Model 2		1	.93
Model 3			1

The correlations of the posterior means for the agenda setting measures across the thee models are (all $p < .001$):

	Model 1	Model 2	Model 3
Model 1	1	.93	.89
Model 2		1	.87
Model 3			1

The correlations are simple assessment (to avoid plotting chains for a visual assessment of 900 parameters) of between-chain agreement. We see that the posterior means from models one and two are most highly correlated. However, I also conduct formal convergence diagnostics.

The Geweke diagnostic tests for equality of the means of the first and last part of a chain. I test the first 50% and last 50% of the turn-level binary topic shift indicators indicated 88.1%, 79.6%, and 85.2% of the 899 topic-shift variable chains passed the test, respectively. Further, using the Gelman diagnostic (Gelman and Rubin 1992; Brooks and Gelman 1998), I find that $R_c < 1.2$ is reached for 93.9% of the 899 topic shift indicators, thus I am fairly confident convergence was reached.

Lastly, 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, 1042 words and 18 topics results in 18,756 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 18 topics across the 3 models is -1449.744, -1436.617, and -1514.061, respectively. Model two has the highest 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 one and two (and vice versa). Likewise, the median number of shared words is 8.5 between the aligned topics of models one and two. This is unsurprising as model two had the best performance on the coherence metric.

Candidate	Reference	Average $L1$	Median number of shared top words
1	2	0.375	8.5
1	3	0.442	7.5
2	1	0.394	8.5
2	3	0.480	7.5
3	1	0.431	7.5
3	2	0.480	7.5

E.3 Topics

Because of the high-dimensional nature of the topic assignment parameters $(z_{d,t,n})$, I retain only every 100^{th} iteration between iterations 1,000,000 and 1,500,000. I then use the modal sampled value as my estimate of each $z_{d,t,n}$.

Because model two has the best performance across the metrics, I present FREX top words from this model. I also present labels for the topics. Labels were chosen from (1) listening to audio recordings and re-reading transcripts of *all* the deliberations, (2) consulting speaking turns and segments highly associated with each topic, and (3) consulting the FREX top words.

Label	Top words
Lack of enforcement mechanisms	who, enforc, more, code, place, syllabus, consist, teacher, stuff, honor
Include photo examples of standards	photo, includ, exampl, pictur, groom, at, suppos, qualiti, show, prefer
Purpose of standards (BYU)	us, we'r, way, case, restrict, ask, whi, societi, concern, which
Writing down proposals	we, do, talk, write, one, this, how, want, down, about
Longterm effects of changing standards	chang, univers, student, then, tri, them, everi, year, idea, want
Ear piercing guidelines	pierc, ear, lds, onc, push, prophet, direct, mormon, two, leav
Reasoning, stating overall opinion	like, that, think, know, they, but, don't, is, it, just
Purpose of standards (church)	church, standard, reflect, understand, read, purpos, guess, command, live, digniti
Women's guidelines	women, leg, definit, fit, feel, cloth, inappropri, sleeveless, wear, guidelin
Personal stories	was, he, came, veri, shave, presid, mission, did, must, didn't
Deliberation guidelines	will, group, the, of, vote, research, to, for, in, inform
Shorts and shoe guidelines	knee, shoe, mean, short, super, flip, walk, my, your, flop
Hairstyles guidelines	extrem, color, style, avoid, purpl, fine, hair, interest, scratch, dye
Five minutes left warning	minut, five, continu, been, begin, you, readi, it, can, here
Personal importance of standards	kindof, as, far, differ, me, appear, care, lot, about, also
Facial hair guidelines	facial, beard, trim, long, clean, neat, should, mustach, hair, well
Punishment for women	she, get, girl, uncomfort, befor, test, run, penalti, turn, work
History of men's grooming in church	their, reason, cultur, bun, beard, religion, grow, except, against, same

F ONLINE DELIBERATIONS

F.1 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.

- 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.

F.2 Chat app and deliberation prompt

When returning for the deliberation, participants were filtered into a chatroom with their pre-assigned partner. Figure 4 shows the Chatter user interface. Deliberation instructions appear at the top of the page. The participant is reminded of the charity they chose and the open-text response they gave in the pre-test survey. Akin to other messaging software, an individual's own messages appear on the right. Their partner's messages appear on the left. When the timer indicates no time is left, the "Done" button activates and redirects users to a post-conversation survey when clicked.

F.3 Text preprocessing

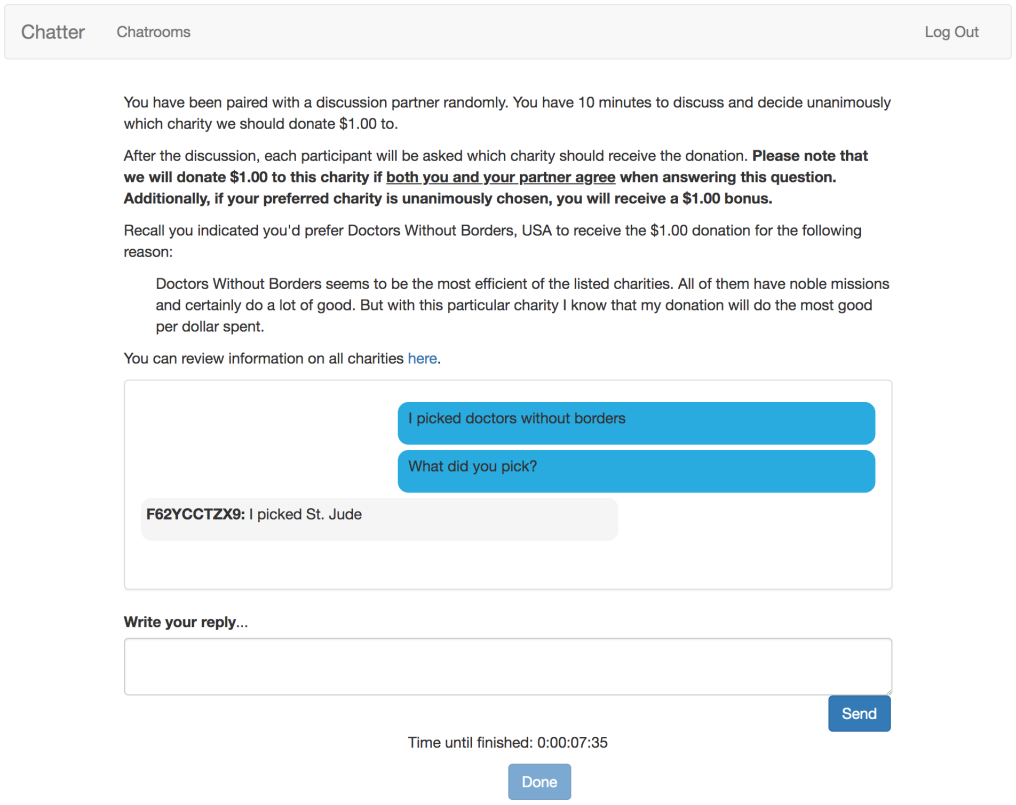
There are 61 deliberations. On average, there are 18 speaking turns per deliberation. To pre-process the text I transformed all words to lowercase and stemmed. I also selectively removed punctuation, keeping exclamation marks, question marks, and punctuation meant to mimic emojis (e.g., ":)"). I kept 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. Finally, I removed all words that did not occur in two or more speaking turns. The corpus vocabulary then contained 913 tokens across 1,098 speaking turns.

F.4 Convergence and topic alignment

I estimated three models with hyperparameters $K = 18$, $\alpha = .01$, $\beta = .015$, $\gamma = 1$ and different starting values. 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 of 1 with probability .05 for the first model, .1 for the second model, and .25 for the third model. The chains ran for 1,500,000. I discard the first 500,000 as burn-in iterations. Moreover, because of the high-dimensional nature of the model with 122 participants, I collect only every 50^{th} iteration. This leaves me with chains of length 10,000 to estimate the posterior distributions of model parameters.

As these are online deliberations, participants express their points and opinions with far fewer words than when verbally communicating. Therefore, I estimate $l_{d,t}$ parameters for all speaking

Figure 4: Chatter user interface and deliberation instructions



Note: Chatter user interface. Instructions appear at the top of the page. Akin to other messaging software, an individual's own messages appear on the right. Other users' messages appear on the left. When the timer indicates no time is left, the "Done" button activates and redirects users to a post-conversation survey when clicked.

turns.

The correlations of the posterior means for the topic shift indicators across the thee models are (all $p < .001$):

	Model 1	Model 2	Model 3
Model 1	1	.97	.98
Model 2		1	.98
Model 3			1

The correlations of the posterior means for the agenda setting measures across the thee models are (all $p < .001$):

The correlations are simple assessment (to avoid plotting chains of over 900 parameters) of between-chain agreement. We see that the posterior means from models one and two are most highly correlated. However, I also conduct formal convergence diagnostics.

The Geweke diagnostic tests for equality of the means of the first and last part of a chain. I test the first 50% and last 50% of the turn-level binary topic shift indicators indicated 90%, 86%, and

	Model 1	Model 2	Model 3
Model 1	1	.97	.98
Model 2		1	.98
Model 3			1

71% of the 1098 topic-shift variable chains passed the test, respectively. Further, using the Gelman diagnostic (Gelman and Rubin 1992; Brooks and Gelman 1998), I find that $R_c < 1.2$ is reached for 94% of the 1098 topic shift indicators, thus I am fairly confident convergence was reached.

Lastly, I do not formally assess convergence of the topic assignment latent variables estimated by the model $(z_{d,t,n})$ for two reasons discussed in Section E.2. Rather, I assess topic coherence and alignment. Topic coherence, when considering the $M = 10$ most probable words for each topic, is consistent across models. The total coherence for all 18 topics across the 3 models is -1717.081, -1698.567, and -1685.145, respectively. Model three best optimizes this coherence metric.

Next, I assess topic alignment, described in Section E.2. There's no clear pattern between the alignment of the models. We see that the median number of shared words between the aligned topics of models one and two is 9. However, the median number of shared words between model three and the others is only slightly less at 8.5 and 8. I report model two's FREX top words below.

Candidate	Reference	Average $L1$	Median number of shared top words
1	2	0.468	9
1	3	0.447	8.5
2	1	0.477	9
2	3	0.465	8
3	1	0.485	8.5
3	2	0.481	8

F.5 Topics

Because of the high-dimensional nature of the topic assignment parameters $(z_{d,t,n})$, I retain only every 100^{th} iteration between iterations 1,000,000 and 1,500,000. I then use the modal sampled value as my estimate of each $z_{d,t,n}$.

Because model two has the highest coherence, I present FREX top words from this model. I also present labels for the topics. Labels were chosen from (1) re-reading transcripts of *all* the deliberations, (2) consulting speaking turns and segments highly associated with each topic, and (3) consulting the FREX top words.

It is worth noting again that I did minimal text preprocessing for the online deliberations. Specifically, I did not remove common stopwords because I found in these deliberations common stopwords or seemingly uninformative parts of speech were particularly informative. For example, I found words like "I've" or "my" were informative of personal experience, whereas words like "them" and "they" were informative of dialogue about how charities help others.

Label	Top words
St. Jude's advertisements	grow, littl, sick, tough, advertis, rather, reput, heartstr, fan, such
Work of American Red Cross	disast, hurrican, never, natur, tornado, flood, impact, kind, caus, wide
Work of Doctors Without Borders	themselv, respect, put, wonder, whole, opinion, doctor, dwb, learn, accur
Importance of donation	dollar, bias, wrong, way, i'll, decent, sens, intern, real, doesn't
Deliberation instructions (beginning)	share, pleas, thank, which, you, particip, by, start, !, chose
Financial health and transparency scores	score, their, most, %, 100, goe, highest, financi, effici, enough
Family/friend/personal with heart disease	heart, aha, die, member, your, famili, continu, altern, someon, problem
Research makes charity worthy of donation	better, etc, benefit, research, worthi, childhood, cure, import, blood, suppli
TV	miss, lol, some, am, cabl, switch, cold, watch, hate, commerci
Family/friend/personal experience with sick child	had, i'v, he, friend, experi, son, big, was, through, my
Stating agreement	decis, agreement, glad, an, two, :), total, togeth, come, understand
Coordinating agreement	minut, 10, agre, fine, suppos, guess, on, want, yes, util
MTurk and HITS	yea, few, cool, dure, hard, time, thousand, just, last, rememb
Concluding deliberation	too, chat, with, nice, have, a, bye, good, great, day
Importance of helping children	they, think, that, but, and, them, children, about, more, all
Access to healthcare	help, peopl, need, care, treat, life, becaus, who, access, medic
Deliberation instructions (concluding)	decid, $, decis, after, unanim, yet, 1, we, should, right
Greeting partner	hello, hi, i, stj, did, ?, donat, whi, chose, choos

F.6 Full regression model

Table 4: Agenda setters more likely to achieve preferred deliberation outcome

	Coefficient	Std. Error
Intercept	0.98	(1.16)
Agenda setting	3.42	(1.95)
Charity-American Heart Association	-0.85	(1.91)
Charity-American Red Cross	-0.76	(1.00)
Charity-Doctors Without Borders, USA	-1.30	(0.76)
Charity-UNICEF USA	1.53	(1.32)
Gender	-0.17	(0.43)
Age-30-39	-0.58	(0.55)
Age-40-49	0.27	(0.71)
Age-50-59	-0.31	(1.10)
Age-60-69	-0.02	(2.23)
Education-Some high school	-1.77	(1.48)
Education-High school diploma or GED	-2.04	(1.10)
Education-Some college	-1.61	(1.05)
Education-Bachelor's degree	-0.90	(0.87)
Education-Master's degree (For example: MA, MBA, MSW)	-0.50	(0.99)
Education-Professional school degree (For example: MD, JD, DDS)	-18.60	(1.42)
Ethnicity-Asian,Pacific Islander,White	-18.02	(1.17)
Ethnicity-Black or African American	-1.07	(1.56)
Ethnicity-Black or African American,White	17.51	(1.50)
Ethnicity-Hispanic/Latinx	1.94	(1.49)
Ethnicity-Hispanic/Latinx,White	-17.53	(1.36)
Ethnicity-Native American	-17.47	(1.34)
Ethnicity-Other	-18.71	(1.47)
Ethnicity-White	0.04	(0.97)
Observations	116	
AIC	177.5	

Note: Coefficients from logistic regressions. Model two reports clustered standard errors at the partnership level in parentheses. Dependent variable is if participant won ($y = 1$) or lost ($y = 0$) the debate. Excluded reference is St. Jude Children's Research Hospital for charity, female for gender, 18 - 29 for age, Associate's degree for education, and Asian for ethnicity.