What Would You Say? Measuring Ideological Speech Patterns with a Novel Survey Instrument

Prepared for the 79th Annual MPSA Conference

William Small Schulz
PhD Candidate, Princeton University
wschulz@princeton.edu

April 5, 2022

Abstract

In this paper, I introduce a novel approach to estimating the ideological positions embodied by political catchphrases prevalent in American political discourse. By taking advantage of the contemporary salience of ideologically-slanted words and phrases, I develop a method to measure the ideological content of ordinary citizens' speech, which is often difficult to observe directly. In overcoming this obstacle, my method offers answers to pressing questions about the representativeness, pliancy, and polarization of mass speech.

Inverting the traditional text analysis research pipeline in which meaning is extracted from natural language documents, I instead devise a novel survey question that generates a data structure identical to conventional text data formats, but which is far more information-dense. This allows me to apply a canonical model of text ideology to not only estimate the ideal points represented by catchphrases and the people who use them (or don't), but also to characterize individuals' overall outspokenness. The behaviors of speaking out and self-censoring (or even falsifying one's preferences) are the subjects of longstanding theories in politics and communications, which I am able to address in an experimental framework by randomly manipulating the social context in which speech hypothetically occurs.

I implement this method in two studies, finding that individuals' speech ideologies can be characterized as a weighted average of concrete issue preferences and abstract political identity, lending some nuance to prevalent debates about "virtue signalling." I also extend past findings regarding the relationships between ideology and demographic traits like gender, age, and education. I characterize the novel trait of outspokenness as a function of news consumption, political interest, and identity strength. I estimate causal effects of social context on speech ideology and outspokenness, finding that individuals are most outspoken amongst close friends, and tend to self-censor political catchphrases from conversations with new acquaintances or political posts on social media platforms, although this effect is moderated by a preference for being more outspoken in more likeminded settings.

Finally, I am able to speak to the issue of polarization in online political discourse, finding that this phenomenon is most likely attributable to extreme speakers' self-selection into participating in political speech online at higher rates than moderate speakers, rather than speakers code-switching into a more extreme speech register in online spaces than they use amongst close friends. This method offers a new approach to study mass speech through the lens of catchphrases, which I find to be quite rich with political meaning.

This research was supported by a grant from the Center for the Study of Democratic Politics at Princeton University.

1 Introduction

Speech is fundamental to politics: it is the medium through which we articulate shared ideals, so that we may pursue them through collective action. While scholars have previously focused on speech by politicians, journalists, and other public elites, today there is increasing interest in speech as a mass political behavior, whose consequences may be greatly amplified by social media. This amplification of mass speech has been attended by an unprecedented degree of attention to the *way* ordinary people talk about politics, including the specific words and phrases we use to express ourselves. Political catchphrases, like "systemic racism," and "Latinx," both reflect and project our agendas and values; they embody the frames through which we understand political issues, and seek to persuade others to share our perspective. Words construct political discourse, and thus offer a strategic tool particularly valuable to constituencies often deprived of the power to set the terms of political debate.

But the usage of such catchphrases also functions as a shibboleth of political identity – an ideological code that grants acceptance within certain communities, and may therefore be highly sensitive to social context. Different situations may induce us to speak out, or to remain silent; in some cases they may even lead us to actively misrepresent our preferences. Indeed, some view such speech as a form of "virtue signaling," employed by liberals as a display of progressive idealism. In this paper I debunk the notion that slanted catchphrases are unique to liberals, but we may still rightly question whether they are anything more than cheap talk. Catchphrases devoid of policy conviction would be vapid affectations, epitomic of a political culture riven by social animosities that belie pervasive indifference to the details of the issues ostensibly motivating political conflict.

These contrasting perspectives imply several well-defined research questions about modern political catchphrases: First, can we identify the discursive ideological positions embodied by catchphrases, and the people who use them on *both* the left and the right? Second, does a person's speech ideology (the ideal point implied by the phrases they use) reflect their concrete issue positions, or merely their abstract ideological social identity? And third, how do people adjust their speech between different social contexts, in ways that might censor or misrepresent their positions?

If mass speech is important, then it is important to answer these questions, but doing so requires an unorthodox methodological approach for estimating the ideological content of mass speech: elite-focused text scaling methods offer a starting point for this measurement challenge, they are ill-suited to answering the questions posed above, because (1) they are not designed for studying words and phrases of *a priori* substantive interest to the researcher, (2) they typically use speech as a substitute for gold standard measures of preferences, rather than as a distinct behavior that might deviate from those preferences, and (3) they rely on convenience samples of documented speech in a single social context, that are generally unavailable for ordinary people. The latter constraint makes it particularly difficult to study self-silencing and abstention from political discourse, which are behaviors just as significant as speaking out – indeed, while social media platforms amplify mass speech and render it observable to researchers, the fact that most users do not talk about politics online means this speech is only a narrow reflection of a talkative few (Wojcik and Hughes, 2019), which both raises important questions about whether and why of online political discourse

is excessively polarized, and simultaneously renders social media data largely useless for answering them.

I therefore introduce a new approach to studying mass speech, which begins with a survey question: I simply ask people whether they would use certain catchphrases, in a specific social context. By experimentally manipulating the social context specified in the question prompt, I am able to test hypotheses about the attitudinal and social origins of political speech, within the familiar framework of a survey experiment. And, because this survey question generates a data matrix structurally identical to – but far denser than – traditional text data, I am able to apply well-established text scaling models (with minor distributional adjustments) to not only estimate individuals' *speech ideology* (the ideal point implied by their self-reported phrase usage) but also their *outspokenness* (their overall propensity to use ideological phrases) on the political topic at hand. Relying only on assumptions common to all survey research, I derive new insights from longstanding text analysis methods, while extending their applications beyond the chattering elite for whom they were originally designed, to include any individual who is willing to complete a questionnaire.

I then apply my method to answer the questions posed above, finding (1) that ideological catchphrases are used by ordinary people on both the Left and the Right, (2) that their usage conveys an overall speech ideology that reflects a weighted average of *both* speakers' concrete issue preferences *and* their abstract social identities, and (3) that social context treatments have robust causal effects on speech, such that individuals are most outspoken when talking to close friends, but self-censor when speaking with new acquaintances and when posting on social media (except when they perceive their online networks to be more likeminded than their close friends). Addressing the aforementioned polarization of online political discourse, I also find that social media users who participate in online political discourse have more polarized baseline speech patterns, compared to people who use social media but don't discuss politics.

The paper proceeds as follows: Section 2, below, motivates my substantive aims in greater detail, and Section 3 explains why past methodological approaches to scaling speech cannot answer the research questions I pose. I then introduce my own method in Section 4, and implement it in two studies described in Section 5. Section 6 presents the results of these studies, and Section 7 concludes by discussing substantive and methodological implications.

2 "Words Matter": The Rise of Political Catchphrases

Speech has long fascinated students of politics, but in recent times the topic has achieved a seemingly unprecedented degree of public interest. The earliest scholars of political communication recognized that interpersonal speech shapes individuals' political attitudes (Katz and Lazarsfeld, 1955), but the rise of social media has greatly amplified the mass public's broadcasting power (Tufekci, 2017): those who speak up online can set the political agenda (Barberá et al., 2019; King, Schneer and White, 2017), and iteratively redefine the norms that structure political coalitions – a discursive power historically reserved for elites (Noel, 2012). Concurrently (and perhaps not coincidentally), a growing movement in the American political landscape asserts that the words and phrases we use in daily life have significant power, and that ordinary citizens should therefore choose their words

with great care.

This attention to word choice seems to have given rise to a burgeoning lexicon of political catchphrases – including the very injunction that "words matter" (McConnell-Ginet, 2020) – which are widely understood to embody a speaker's worldview, like "Black lives matter" or "America first," denote their orientation towards a person or group, like "thug," "snowflake," or "patriot," or evoke a particular framing of a topic, like "China virus," or "systemic racism." This ideological coding of language has become so extensive as to provoke some consternation amongst everyday political speakers who are anxious to express themselves in the correct terms – what McWhorter (2021) describes as "lexicographic paranoia." I proceed to ask three questions about this phenomenon:

1. Are political catchphrases peculiar to liberals, or do conservatives also choose their words in ways that reveal their ideological positions?

2. Do these words and phrases have concrete policy content, or are they merely signals of speakers' political identities (so-called "virtue signalling")?

3. Do people use this language consistently across different social contexts, or do they engage in ideological code-switching? In particular, how does speech with friends differ from speech with strangers, and how does online discourse differ from traditional conversation?

In the remainder of this section, I elaborate these questions, before turning to discuss methods for answering them.

2.1 Ideological Lexicons: Progressive Quirk or General Phenomenon?

The casual observer of American politics might easily conclude that attention to language is a distinctively left-leaning behavior. Certainly, contemporary liberal discourse is awash in neologisms designed to respect marginalized groups, like "Latinx," and construct as harmful certain behaviors that might otherwise seem innocuous, such as "mansplain" and "micro-aggression" (Harmon, 2021) – coinages oft-derided by conservatives as "wokespeak" (Hanson, 2020). Progressive arguments for regulating language are widespread in the popular press (e.g. DiAngelo, 2021), and endorsed by major institutions such as the American Medical Association, whose 2021 "Language Equity Guide" instructs health professionals on "...disavowing words that are rooted in systems of power that reinforce discrimination and exclusion," (AMA) and replacing them with preferred alternatives, such as "undocumented immigrant" instead of "illegal immigrant," and "oppressed" instead of "vulnerable."

To be sure, there are also left-of-center skeptics who reply that "...this kind of linguistic fussing matters less than whether or not we tackle the material roots of deprivation and inequality." (Yglesias, 2021) Nonetheless, the progressive lexicon has reached a degree of salience that arguably meets the linguistic definition of a *sociolect*: a way of speaking associated with a particular social group – in this case, American liberals. Viewed this way, progressives in the United States are members of a *linguistic community*, defined by a set of speech norms that constitute a *language ideology*[1] that

[1]The linguistic definition (Errington, 1999) of a "language ideology" is quite broad, typically referring to hegemonic belief systems that map language variations used by ethnic, economic, and other social groups to the hierarchical social status valuations of those groups. However, I use the term in a political sense throughout this paper, since I am applying the concept to describe a mapping between lexical variation and political orientation.

"privilege[s] certain ways of speaking as inherently 'better' than others" (Wardhaugh and Fuller, 2015, p. 75). Usage of these terms thus functions as an *index* (Silverstein, 2003) of a speaker's social identification as a liberal (c.f. Mason, 2018).

Yet the Right can be equally fussy about words: "The struggle over the lexicon is actually the central struggle," according to Stephen Miller (Shear, 2021), speechwriter to Donald Trump, whose administration banned the Centers for Disease Control from using a list of terms including "vulnerable,"[2] "diversity," and "transgender," (Sun and Eilperin, 2017), stripped government websites of references to "climate change" and "clean energy" (Rinberg et al., 2018), and instructed US attorneys that they must use the phrase "illegal alien" and not "undocumented immigrant" to refer to persons present in the US without authorization (Kopan, 2018).

From these examples, it might appear that the Right's linguistic fussing is primarily elite-driven. We might expect rank-and-file conservatives, who identify with the plainspokenness of "telling it like it is" (or more colorfully, the ethos embodied by the phrase, "fuck your feelings"), to give relatively little thought to the specific words they use. The Right's attention to language may be, in part, a reaction against the Left's language activism, but this does not make it any less ideological: conservatives' avoidance of liberal-coded terms is a manifest behavior that is equally rich with political meaning as progressives' gravitation towards them – both behaviors reveal a shared ideological understanding of language. Moreover, as I demonstrate in this paper, it is possible to identify many conservative-coded phrases (not unlike those already cited in this paragraph) that are affirmatively used by people right-of-center. I present these empirical findings at the beginning of Section 6, but I hold the reader in no suspense on this point: this paper finds that ideological catchphrases span the political spectrum.

2.2 Virtue Signaling: Do People Mean What They Say?

If political catchphrases are a prevalent feature of American political speech, we might ask what, precisely, they signify: do they have policy content, or are they merely signals of identity? The liberal lexicon in particular is often criticised as a form of "virtue signaling" (c.f. Bartholomew, 2015), more performative than substantive. Performances can play an important role in constructing and revising political identities (Cramer, 2004; c.f. Butler, 1999), but if speakers are simply conforming to the dominant terminology in their speech communities (Pagel et al., 2019), the evolution of this lexicon may reflect a state of pluralistic ignorance (Bicchieri, 2005), such that members of these identity groups espouse views they do not actually hold (Kuran, 1995).

Coaston (2017) argues the "virtue signaling" framing is pernicious, in that it encourages us to doubt whether others truly mean what they say, simply because they claim to care about issues that we do not. Yet the ambiguity is very real: for example, the slogans "defund the police" and "abolish the police" have become central themes of contemporary American policing discourse (Yglesias, 2020), despite genuine disagreement about whether these constitute demands for European-style social services to displace police from roles that need not involve violence (Lowrey, 2020) or the literal abolition of policing as an institution (Kaba, 2020). This echoes similar ambiguities on the

[2]This term seems to be at once too progressive for the Trump Administration, and too reactionary for the AMA, such that a person wishing to discuss this topic might reasonably wonder whether any term could satisfy both sides.

Right, including the 2016-era debate about whether to take then-candidate Donald Trump's statements "literally" or "seriously" (Zito, 2016), as well as ongoing doubts about whether his supporters truly believe "The Big Lie" that he rightfully won the 2020 election, or if this might simply reflect expressive responding (Arceneaux and Truex, 2021; Cuthbert and Theodoridis, 2022).

Thus, we are left to question whether the words people use are actually representative of their opinions, on either side of the aisle. It is plausible that this kind of speech reflects an "Overton window" strategy of exaggerating one's demands for persuasive effect (Simonovits, 2017), but it may also create a one-way ratchet that emboldens our language while leaving the underlying meaning unchanged – a phenomenon known as *semantic bleaching* (Traugott, 2006). I therefore pose the following research question: Is an individual's usage of political catchphrases reflective of their concrete policy preferences, independent of their ideological identity strength?

2.3 Context Effects: Ideological Code-Switching and Self-Censorship

Finally, in addition to examining the *content* of political catchphrases, I investigate the role of social *context* in shaping how people speak, which is a subject of longstanding interest in the communications literature. In particular, Noelle-Neumann's (1974; 1991) "Spiral of Silence" theorizes that people have a "quasi-statistical sense" of the distribution of public opinion, and use this knowledge to self-censor opinions they perceive to be unpopular when speaking in public – a scenario often operationalized as a hypothetical conversation with a stranger on a train. Given the potentially loose linkage between speech and preferences discussed above, individuals might go further than self-silencing, and "falsify" their true preferences (Kuran, 1995), in the interest of making a positive impression on their immediate audience (c.f. Goffman, 1956; Turner, 1991).

This line of thinking has implications for political discourse on social media platforms: in particular, it suggests that online polarization could be illusory, as exemplified in popular debates about whether or not "Twitter is real life" (Goldberg, 2019; Warzel, 2020). Descriptive evidence shows that US Twitter users skew unrepresentatively Democratic, and that 80% of tweets are authored by the most prolific 10% of Twitter users, who are disproportionately inclined to tweet about politics relative to the average user (Wojcik and Hughes, 2019). This pertains to a broader scholarly concern that the political speech that occurs on social media platforms is unrepresentative and contributing to polarization, by exposing users preferentially to attitude-reinforcing content (Sunstein, 2017), or distasteful out-group discourse (Settle, 2018) – processes that would be exacerbated if the way people talk online is unrepresentatively polarized in the first place.

Some argue that the apparent extremity of online political discourse is merely a consequence of hostile individuals self-selecting into online political expression at higher rates (Bor and Petersen, 2021). Others highlight the role of speech norms, showing that users' expression of moral outrage online can be amplified by a process of social learning (Brady et al., 2021), and that toxic speech can have harmful downstream effects in discussion threads (Guess et al., 2019). To borrow another concept from linguistics, we might speculate that hostile online speech norms could induce users to engage in *code-switching*, adopting more extreme ideological language online than they would use offline. I thus pose a final research question: Is online discourse characterized by polarized speech

because users adopt more extreme speech patterns online than they use with close friends, or because the people who self-select into online political expression are more extreme speakers than those who remain silent?

3 Methodological Background: Text Ideal Point Models

To answer these questions, we might be tempted to apply one of the various methods furnished by the political science literature to transform text data into ideal points. However, the existing methods cannot answer the questions posed above, because (1) they are not designed to treat words themselves as important units of analysis (and thus are ill-suited to shed light on the lexico-political trends that motivate the present study), (2) they are intended to produce ideal points as substitutes for gold-standard measures of "true preferences" (and thus can't identify whether speech deviates from those preferences), and (3) they generally rely on speech derived from a convenience dataset obtained in a single context, such as congressional floor speeches, social media posts, or party manifestos (and thus cannot characterize people who did not speak in that context, nor measure how individuals might code-switch between different contexts).

These limitations are, perhaps, a natural consequence of the research goals that motivated the original development of text scaling methods: Laver, Benoit and Garry (2003) first introduced the Wordscores method of text scaling with the stated goal of reducing the effort required to infer parties' ideological positions based on their published manifestos (these estimates being useful for subsequent comparative analyses). The substantive meaning of the words themselves was merely incidental: in applying their method to scale German political parties "using no knowledge of German," (p. 325) the authors conclude, "our technique allowed us to analyze very quickly and effectively texts written in a language that we do not speak!" (p. 326) So, although Wordscores represented a significant methodological innovation for relating speech to ideology, it was not designed to study lexical features of *a priori* substantive interest to the researcher.

Although text analysis methods have advanced in many ways since the original publication of Wordscores, most studies display the same merely-incidental interest in text as a substitute for missing information, and therefore have similar limitations. Slapin and Proksch (2008) introduced the Wordfish method of text scaling (which I discuss further below, when I introduce my own method) as an improvement upon Wordscores, while maintaining the same goal of measuring parties' policy positions as a prelude to comparative analyses. Gentzkow and Shapiro (2010) famously used text scaling to link legislators and media outlets, so that the ideal points assumed of the former could be used to infer the political biases of the latter. While they report several findings relevant to the present study – such as that Democrats use the phrase "estate tax" while Republicans say "death tax" – these are a sideshow to the main attraction of creating an ideological scaling of media outlets by taking advantage of the fact legislators and journalists both[3] use words. As aptly summarized by Monroe, Colaresi and Quinn (2008), when such papers report the ideological weights of specific

[3]However, their identification strategy requires that they use these words similarly, despite the profoundly different substantive contexts of congressional speech *versus* journalism – in this paper I introduce a method to measure the differences between contexts.

words used in their analyses, "they are often intended (implicitly) to offer semantic validity to an automated content analysis," (p. 373) rather than as an object of interest in their own right.

The advent of modern social media has made mass speech available for similar analyses, but most of these studies continue to use text as a solution to a missing-data problem: Most frequently, scholars seek to use social media text to measure public opinion in the absence of polling data (Klašnja et al., 2015). As stand-ins, such measures are typically optimized for predicting attitudes and traits as measured by surveys or other external gold-standards: Barbera (2016), for example, trains classifiers to predict sociodemographic traits from publicly available data on Twitter users. Others use text analysis to estimate quantities that would otherwise be difficult to measure, such as area-level racism (Nguyen et al., 2020). A few papers use hand-labeled text data to train classifiers to measure quantities that are distinctive to speech itself, such as deliberative quality (Jaidka, 2021), toxicity (Guess et al., 2019), and moral-emotional language (Brady et al., 2017). In closely related work (Schulz et al., 2020), my colleagues and I apply similar supervised learning techniques to measure liberal-conservative signals in tweets. However, these approaches remain bound to a single social context in which convenience data are observed, and so can't compare how people speak in different contexts.

Moreover, supervised learning with an open vocabulary is generally unsuitable to understanding lexical differentiation as a social phenomenon: while the optimal feature set for a prediction task may include words and phrases that authors consciously use to signal their preferences and identities, these procedures may also place significant weight on features that reflect subconscious psychological traits with no explicit social meaning (e.g. Schwartz et al., 2013), or terms that reflect topical differences (Roberts et al., 2014) in what people want to talk *about*, rather than *how* they want to talk. Such open-vocabulary methods are thus an inappropriate tool for measuring language ideology, which arises from the political connotations that speakers explicitly know and use to guide their usage of specific words and phrases that are salient in the broader political discourse. Indeed, contemporary state-of-the-art text analysis methods transform text into an abstract vector space (that better represents documents' semantic meaning) before learning prediction weights, which improves classifiers' performance (e.g. Devlin et al., 2019) but produces no insight at all into the systems of shared meaning that structure ideologues' choices to use (or not use) specific human-readable words or phrases, which is the object of my study.

Thus, existing text analysis methods lack several features that are necessary for my study: they do not measure the substantive meaning of words, they equate text-based measures of ideology with policy preferences, and they use convenience data observed in a single context. This presents a serious problem, since I wish to study the role that specific words and phrases with known political connotations play in constituting a linguistic ideology that may deviate from policy preferences, and which may vary in consequential ways between different social contexts in which a person might speak.

4 A New Approach: What Would You Say?

My solution is quite straightforward: rather than observing what people say in a messy convenience dataset drawn from a single context, I collect clean and focused data using a specially-designed questionnaire, which simply asks people what they would say, in a specified social context. An example of this "What Would You Say?" (WWYS) question can be seen in Figure 1. Participants are presented with a grid of "words and phrases that someone might use when talking about politics," and are asked to report their propensity to use each word or phrase on a three-point scale ("I would say this"/"I might say this"/"I wouldn't say this"), in a social context that can be manipulated to test hypotheses about ideological code-switching and self-censorship. This approach has several virtues:

First, it allows me to study specific words and phrases that are of *a priori* substantive interest. I developed the list of words and phrases included in my instrument by applying substantive knowledge of politics, as well as insights from computational text analyses, to create a list of words and phrases that I expected to have widely-known political connotations. This process included the application of a lasso-regularized binomal regression model to select terms and phrases used significantly more often by liberal or conservative Twitter users (see Appendix I Haven't Written Yet), using data from a parallel project (Schulz et al., 2020). However, the final selection of terms was not limited to those identified by this exercise; rather, my goal was to include terms covering a variety of substantive topic areas such as race, gender, nationalism, immigration, policing, and others suggested by colleagues and friends. Further to my expectation that language is linked to ideology not only on the Left, but also on the Right, I took care to include phrases, like "patriot" and "reverse racism" that embody a conservative orientation. Above all, though, I sought to capitalize fully on the advantages afforded by my method, to study words and phrases that ordinary people find interesting and meaningful.

Second, my survey instrument allows me to manipulate the social context of speech while holding all other variables constant. In the studies reported below, I randomly assign participants to various social context treatments, including conversations with close friends, encounters with new acquaintances, and posting to online platforms. As discussed below, I recover significant and robust treatment effects from these manipulations, which not only shed light on the social motivations of political speech, but also suggest that this framework may be extended to study speech in any social context that might be of interest to the researcher.

Third, the responses to this question format can be arranged into a data matrix Y of dimension $I \times J$, where rows represent people, columns represent phrases, and cell $y_{i,j} \in \{1, 2, 3\}$ contains the ordinal value of the response given by person i to phrase j, on the scale *wouldn't* (1), *might*, (2) or *would* (3). This matrix is structurally identical to the document-term matrices (DTMs) on which text models are traditionally estimated. However, naturally-occurring DTMs (those made from text collected "in the wild," so to speak) are usually populated with problematically sparse and noisey data, drawn from extremely skewed count distributions (Lowe, 2001, c.f. Zipf 1949) whose rate parameters have many unobserved sources of variance. By contrast, the cells of the DTMs generated by the WWYS question are drawn from a well-defined ordinal distribution, and are thus far denser than an natural DTM of comparable dimension could possibly be, and are also

Here is a list of words and phrases that someone might use when talking about politics, either online or in a face-to-face conversation.

Please indicate whether each word/phrase is something **you would use with a close friend, who knows you very well**.

	I Would Say This	I Might Say This	I Would Not Say This
"...systemic racism..."	○	○	○
"...MAGA..."	○	○	○
"...big government..."	○	○	○
"...wear a mask..."	○	○	○
"...human rights..."	○	○	○
"...America first..."	○	○	○
"...blue lives matter..."	○	○	○

	I Would Say This	I Might Say This	I Would Not Say This
"...white trash..."	○	○	○
"...thug..."	○	○	○
"...toxic..."	○	○	○
"...post-truth..."	○	○	○
"...mainstream media..."	○	○	○
"...woke..."	○	○	○
"...defund the police..."	○	○	○

	I Would Say This	I Might Say This	I Would Not Say This
"...latino..."	○	○	○
"...abolish the police..."	○	○	○
"...right to work..."	○	○	○
"...all lives matter..."	○	○	○
"...undocumented immigrant..."	○	○	○
"...micro-aggression..."	○	○	○

I Would Say This	I Might Say This	I Would Not Say This

Figure 1: Speech ideology question, under the "close friend" treatment condition. In study 1, the alternative condition asks about phrase usage "...with someone you just met, who doesn't know you too well." In study 2, the alternative condition asks about phrase usage "... on [platform_name]," where the name of the platform selected by the respondent as their primary venue of online opinion expression is piped in for [platform_name].

more accurate (if we believe that survey responses are accurate in general) since I hold constant all aspects of the data-generating process other than the speech behavior the respondent wants to report, and the way the respondent interprets the response format.

I thus invert the traditional text analysis research algorithm of reducing convenience documents to a simplified data structure from which to extract an extrinsically-valuable quantity using an ad-hoc model, and instead start with a canonical model (see Section 4.1, below), to estimate textual ideology for its own sake, and develop a survey question to generate data in the native format required by that model. Although Roberts et al. (2014) introduce methods for analyzing free-text responses collected in surveys, my method skips the step where respondents write actual text (which is an inconveniently high-variance outcome). In essence, I ask respondents to behave as though the "bag of words" model of text, which Grimmer and Stewart (2013) describe as a "shocking" (p. 272) simplification of texts' data generating process, is literally true: I ask them whether they would let certain phrases come out of their mouths. I justify this on the basis that in the context of this study, people in a sense truly are "bags of words" that signal their political orientation independent of syntactic[4] context. In many respects, this approach seems to out-perform convenience text analysis: I am able to apply a conventional text model and extract extensive information from a relatively small-dimension dataset – as discussed below, I even observe robust treatment effects on latent traits that previous scholars have deemed nuisance parameters with little substantive meaning.

4.1 Wordsticks: An Ordinal Extension of Wordfish

I estimate[5] the following model to infer an ideological scaling of each person i and phrase j from the data matrix $Y^{I\times J}$:

$$\Pr(y_{ij} \geq k) = \frac{\exp(\mu_{ij}^k)}{1+\exp(\mu_{ij}^k)}; \qquad \mu_{ij}^k = (\alpha_i - c_j^k) \times \gamma_j + \beta_i$$

Where

$y_{ij} \in \{1,2,3\}$ the observed ordinal response to phrase j reported by respondent i

$k \in \{2,3\}$ the possible response categories (excluding 1, the lowest category)

α_i the speech ideal point of respondent i

β_i the outspokenness of respondent i, reflecting propensity to say phrases, net of ideology

c_j^k the ideology cutpoint between response $k-1$ and k for phrase j

γ_j the ideological slant of phrase j

This model is nearly identical to Wordfish, as specified by Slapin and Proksch (2008) to scale party manifestos (see proof in Appendix A), except that where these authors used a Poisson distribution (with rate parameter equal to $\exp(\mu_{ij})$) to model word counts observed in natural documents, I use an ordinal logistic distribution to model self-reported phrase usage on a 3-point scale that, for each phrase, segments the ideological spectrum at response cutpoints c_j^k (which replace Wordfish's

[4]Though not necessarily pragmatic context, as I effectively manipulate aspects of pragmatics via the social context treatments.

[5]Further details are included in Appendix A, and the model code can be found in Appendix B.

word-level intercept parameter). While nearly all results presented below can be replicated using the original Wordfish model (see Appendix I), this ordinal adaptation is more appropriate to the actual response format, and lends itself to intuitive visualization of phrase ideologies as segments of the ideological spectrum in which each of the 3 possible responses is most probable (see Figure 3). I therefore dub my model "Wordsticks" (see also Goplerud, 2019), in homage to Wordfish, and in reference to this visual intuition.

Note, however, that while Slapin and Proksch (2008) specify a document-level intercept that is mathematically identical to my β_i, they treat it as a nuisance parameter that is needed simply to "control for the possibility that some parties ... have written a much longer manifesto" (Slapin and Proksch, 2008, p. 709). Due to the ordinal structure of my data, however, I am able to substantively interpret this parameter as respondents' *outspokenness*: their overall propensity to affirmatively use phrases that express their ideological position, independent of what this position is.[6] As discussed above and demonstrated empirically in Section 6, outspokenness is a key dimension of political speech orthogonal to ideology, and it is important to measure and explain its variation across members of the mass public, and between different social contexts in which the same individual might speak out, or remain silent. The ability to characterize the relative stridency of speech is a key feature of my method, made possible by the uniquely dense and clean text data generated by the "What Would You Say?" question.

Wordsticks is also a close cousin to the spatial choice models commonly used in legislator ideal point scaling (e.g., Poole and Rosenthal, 1985), with the caveat that many of these models use a symmetric loss function to model votes as a function of proximity between legislator and legislation ideal points, whereas I model catchphrase usage as monotonically[7] increasing or decreasing (depending on the sign of the γ phrase slant parameter) in respondent speech ideology, as is conventional in Item Response Theory approaches to ideal point estimation (c.f. Clinton, Jackman and Rivers, 2004). Except for this difference, Wordsticks is also nearly identical to Barberá's 2015 "Tweetscores" model, which infers Twitter users' ideal points from the accounts they follow. Barberá includes the same β person intercept, which he interprets as "political interest," which is theoretically and empirically consistent with my substantive interpretation of β as outspokenness, in the context of my study of speech.

So, this study not only introduces a novel data collection approach to scale the spoken ideologies of members of the mass public using well-established text analysis models; it also extends previous modeling approaches to estimate a substantively-important dimension of speech: outspokenness.

5 Implementation

I now describe two studies in which I applied my method, to answer the research questions posed above. In both studies, each participant responded to 20 phrases that were sampled[8] from a larger

[6] My inclusion of the intermediate "might say this" response point is also useful for recovering information about outspokenness, as I demonstrate in Appendix J

[7] I interrogate the monotonicity assumption in Appendix H, and find it satisfied for the vast majority of phrases.

[8] The first six phrases (as seen in Figure 1) were held constant, to ensure a sufficient overlap of phrases answered by each respondent. These phrases were selected with the goal of being relatively salient and clear in their political mean-

superset of 46 phrases, so that each respondent responded to a cognitively manageable[9] number of items, while still allowing each study as a whole to cover diverse aspects of the American political lexicon. The results reported below are based on the respondent and phrase parameter estimates derived from estimating the Wordsticks model on the pooled responses from both studies 1 and 2, although in exploring the predictors of these traits I generally estimate separate regression models for each study, due to differences in covariates measured and experimental treatments applied in the two studies.

Study 1 was fielded on Prolific from April 9-10 2021 with a sample of 503 US adults, using quotas to ensure a roughly equal proportion of liberals, moderates, and conservatives. This study manipulated the social context specified in the WWYS question so that respondents imagined speaking either "with a close friend, who knows you very well," or "with someone you just met, who doesn't know you too well," in a between-subjects randomized design. By comparing speech with close friends to speech with strangers, I sought to address longstanding theories about self-censorship (e.g. Noelle-Neumann, 1974) and preference falsification (e.g. Kuran, 1995) in public political speech.

Study 2 was fielded on Amazon Mechanical Turk from September 17-19 2021, with a sample size of 798 US adult participants, and focused on social media users, recruiting only participants who indicated that they were users of either Twitter or Facebook (in addition, the same ideological quotas were used as in Study 1). In this study, participants were first asked whether they used any of their social media accounts[10] to post their views on politics or current events; those who said they did (the *posters*) were subsequently randomized into one of two social context treatments: one was identical to the "close friend" condition of Study 1, while the other asked what language they would use "on [platform]" where [platform] was replaced with the name of the social media platform the respondent nominated as their primary venue of online political expression. Those who indicated they did not use any of their social media accounts for expressing their political views (the *lurkers*, who may encounter political discourse online but abstain from participating in it) received only[11] the "close friend" context treatment. Applying these treatments to the lurkers and posters allowed me to measure two quantities of interest for explaining the polarized speech environments of social media platforms: First, by measuring descriptive differences between the close-friend speech ideologies of lurkers *versus* posters, I can test the self-selection hypothesis that online discourse is polarized because the posters have more extreme baseline speech patterns than the lurkers (by using the close-friend condition as a common reference point). Second, by estimating the causal effect of the social media treatment (relative to the close friend treatment), I can test whether online platforms induce users to code-switch and speak in more extreme ways than they do in traditional offline conversations with friends. Additionally, I asked the posters whether their online networks and close friends were more liberal than them, more conservative than them, or similar to them.

ing, to reduce heterogeneity in participants' interpretation of subsequent items. This also has the effect of weighting the estimation of phrase parameters more strongly to the usage patterns for these clearest phrases, which clarifies the substantive interpretation of the phrase scalings, and makes model identification somewhat more convenient.

[9] Median completion time for 20 phrases was 51 seconds.

[10] Although recruitment targeted Facebook and Twitter users, respondents were asked about their usage of a variety of platforms, described in detail in Appendix I Haven't Written Yet 3

[11] Because these individuals said they did not express their political views online, I did not expose them to the social media context treatment, because it would be difficult to interpret whether their responses would reflect a hypothetical scenario in which they *did* talk about politics online, or the realistic scenario in which they do not do so.

6 Results

I begin this presentation of results with a visual overview of the model parameters and the manifest data from which they were estimated. Figure 2 illustrates the relationship between observed phrase usage and the two latent traits estimated for each respondent, speech ideology α (x axes) and outspokenness β (y axes), for two exemplary phrases: "systemic racism" (left-slanted) and "America first" (right-slanted). The position of each point represents each person's estimated location in the 2-dimensional space of ideology and outspokenness, and so the distribution of points is the same for both phrases: a funnel or diamond shape, with no observations in the corners. The model thus encodes the substantive assumption that no strong ideologues may say all or none of the phrases, since it is *selectivity* in phrase usage that reveals a liberal or conservative position in the speech space. Appendix F provides plots like these for all 46 phrases.

The color of each point, meanwhile, represents each person's observed response to the phrase in question, where green, yellow, and red points denote "would," "might," and "wouldn't say this" responses, respectively. So, we can perceive the ideological slant of these phrases by observing the orientation of the color gradient in these plots: "systemic racism" is a left-slanted phrase, in that participants become more likely to say this phrase the further left is their speech ideal point. "America first" shows the opposite pattern, reflecting its rightward slant. This also illustrates how avoiding a phrase can be just as meaningful as using it: conservatives' refusal to say "systemic racism" is no less ideological than their affirmative usage of "America first." Meanwhile, the up-down color gradients reflect orthogonal variation in outspokenness, such that some people are more inclined to use these kinds of political catchphrases, independent of their left-right ideological position in the speech space.

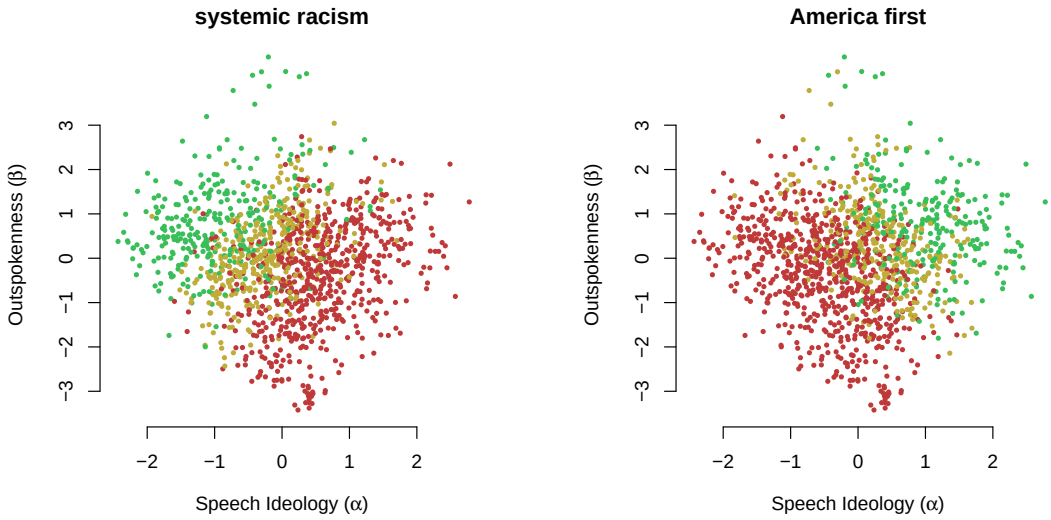

Figure 2: Observed responses to a liberal phrase (systemic racism) and a conservative phrase (America first), plotted by respondent latent traits α (ideology, x axes) and β (outspokenness, y axes). Green, yellow, and red points denote "Would," "Might," and "Wouldn't" responses, respectively.

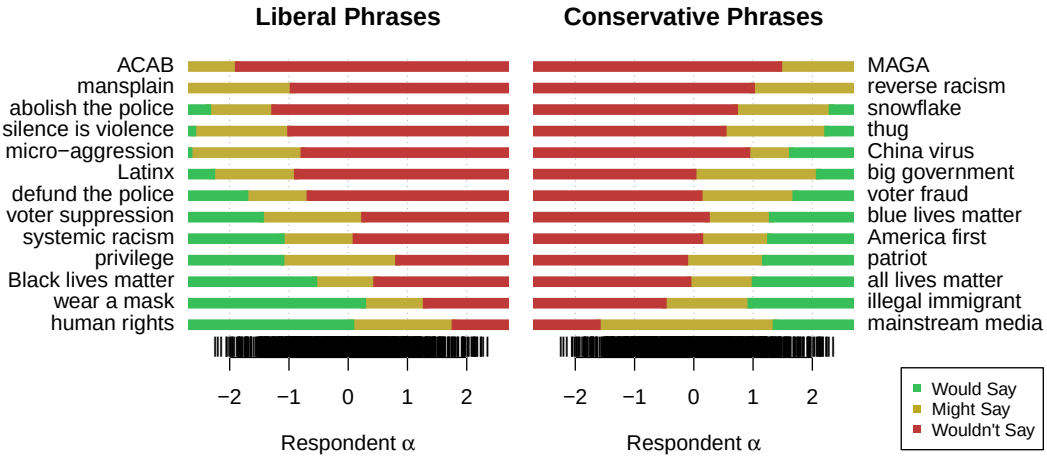

Figure 3: Ideological scaling of selected phrases, in terms of predicted response at different points of the speech ideology (α) spectrum. Green, yellow, and red bands denote "Would," "Might," and "Wouldn't" responses, respectively. Estimated respondent α ideal points are plotted as ink densities above the x axes.

For a more parsimonious ideological scaling of the phrases, Figure 3 plots the predicted response category for each[12] phrase as a function of respondent position along the α ideology dimension (fixing respondent β outspokenness) and the estimated phrase cutpoints $c_j^{2,3}$, and arranges liberal and conservative phrases according to the midpoint of the yellow "might" region as a proxy for the phrase's ideological extremity. This offers semantic validity to the analysis: for example, I find that "abolish the police" is more leftist than "defund the police," and that "blue lives matter" is to the right of "all lives matter". More importantly, though, it supports a central claim motivating this study, that contrary to common perceptions that the American ideological lexicon is a distinctively liberal invention, there are many right-slanted phrases like "snowflake," "thug," and "patriot," that conservatives actively use to convey their positions, just as liberals do via "mansplain," "micro-aggression," and "privilege." Political catchphrases appear to be common on both the right and the left, and their usage can be well-summarized by a single latent dimension of speech ideology.

But what is speech ideology – a statement of issue positions, or of social identity? Figure 4 illustrates that speech ideology correlates strongly with *both* 1-dimensional issue ideology, constructed from 10 agree-disagree issue preference questions (listed in Appendix K) using principal components analysis, *and* ideological identity strength, constructed from 3 items (adapted from Huddy, Mason and Aarøe 2015, see Appendix L) likewise scaled together using PCA. Based on these scatterplots, speech ideology seems to track quite closely with issue ideology, and somewhat more noisily with ideological identity. However, as illustrated in the coefficient plot in the rightmost panel of Figure

[12]Note that this is only a subset of phrases included in these studies, selected for their topical breadth and strong slant. The weaker a phrases' γ slant is (in absolute value), the less informative a respondent's ideological position is in predicting whether they will say it, so this visualization approach is most useful for the most slanted phrases – those that load most strongly onto the dimension of speech ideology estimated here. Appendix G offers a version of this plot that includes all 46 phrases and visualizes their discrimination through color saturation. Appendix E visualizes the γ slant/discrimination parameter estimates directly.

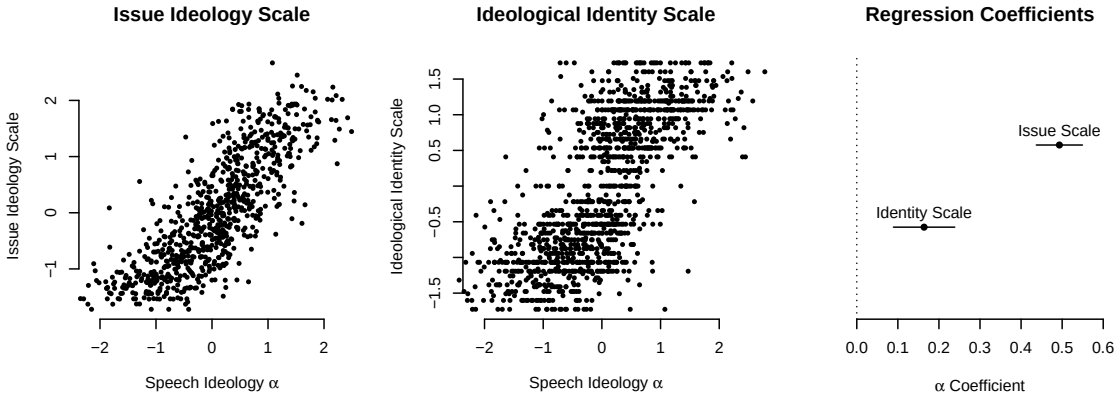

Figure 4: Comparison of speech ideology estimates (y axis) against 2 alternative measures of ideology (x axis): issue ideology derived from 1-dimensional scaling of 10 issue preference items (left panel), and signed identity strength constructed by 1-dimensional scaling of 3 ideological identity questions and interacting with a signed indicator for liberal/conservative identification (center panel). The right panel plots regression coefficients for both of these measures, in a model with respondent speech ideology α as the dependent variable (see Table 10, column 3).

4, both are independently predictive of speech ideology. From their relative magnitudes, we might conclude that speech is predominantly a reflection of one's issue positions – a normatively desirable relationship, if we view speech as a mode of democratic self-representation – but that it also reflects ideological identity. This is consistent with a view of political speech as a behavior that is partly an articulation of policy preferences, and partly an expression of raw political identity, independent of policy preferences, which may be less normatively desirable.

What else can we glean regarding the nature and origins of political catchphrase usage? Continuing in a linear regression framework, Figure 5 presents descriptive analyses identifying key political and demographic predictors of speech ideology α and outspokenness β, and estimates of the causal effects of the social context treatments on both of these dimensions.

First, in the left panel of Figure 5, we can see that speech ideology α correlates strongly with the conventional likert-scale measures[13] of liberal-conservative ideology (as it also did with the issue and identity measures shown above) and partisanship. We also see several significant relationships with demographic variables: First, college education is associated with more left-leaning speech, which perhaps reflects that colleges are generally left-leaning sites of political socialization, and that the left-slanted catchphrases included in this study (like "micro-agression" and "systemic racism") are generally of a more intellectual or academic nature than their right-slanted counterparts. Meanwhile, male gender (as compared to female or non-binary gender) is associated with more right-leaning speech – this would seem to be an extension of past evidence of gender differ-

[13]Both ideology and partisanship are here measured with the same question structure, in which those who choose "moderate" ideology or "neither" party are asked whether they lean one way or the other, generating a 6-point scale for both quantities. Other measures of ideology (namely issues and identity) were excluded from this model, to aid interpretation of the coefficients on these likert-scale measures of ideology and partisanship.

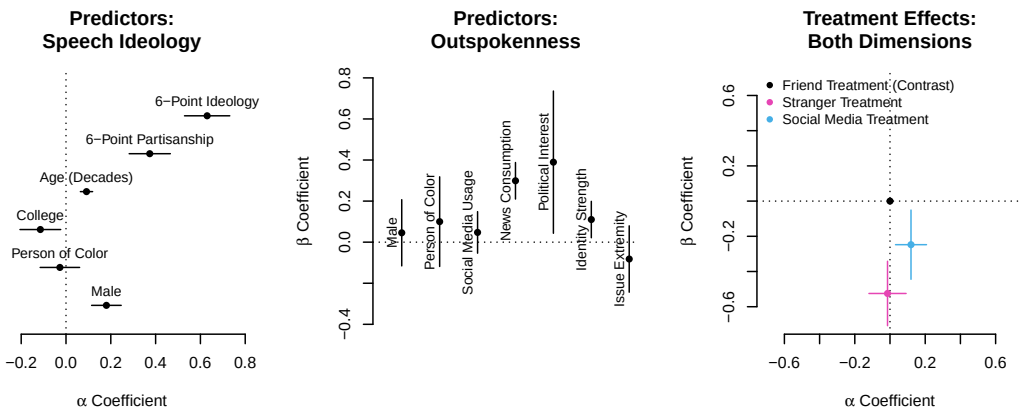

Figure 5: The left panel plots coefficients, with 95% confidence intervals, from a regression model with respondent speech ideology α as the dependent variable (see Table 10, column 5), using pooled data from both studies. The center panel plots coefficients, with 95% confidence intervals, from a model with respondent outspokenness β as the dependent variable (see Table 11, column 7). The right panel plots the effects of the "stranger" (pink) and "social media" (blue) treatments, relative to the common contrast of "close friend" (black, by construction at the origin) in both dimensions: the x axis denotes left-right shifts in speech ideology α (see Table 12, columns 2 and 4), and the y axis denotes up-down shifts in outspokenness β (see Table 13, columns 2 and 4).

ences (orthogonal to traditional concepts of ideology and partisanship) in political speech (Schwartz et al. 2013; Monroe, Colaresi and Quinn 2008, c.f. Gilligan 1993). Older individuals speak significantly more conservatively than younger ones (although the scale of this variable makes it difficult to see), which is consistent with longstanding evidence that age is associated with conservative issue positions – although it is worth emphasizing that the association between age and right-leaning speech is independent of issue ideology (see Table 10, model 3), and so may also reflect that the Left often coins catchphrases for the specific reason of replacing "old-fashioned" terminologies with more progressive alternatives that may be unfamiliar to older people. Finally, identifying as a person of color (as opposed to white) shows no significant relationship with speech ideology, which is perhaps surprising, given the preponderance of race- and ethnicity-related terms included in this study.

The center panel of Figure 5 presents a similar analysis of outspokenness β, which is orthogonal to ideology, but significantly predicted by news consumption (using a summary index spanning 6 types of media, which are analyzed in disaggregated form in Table 5, see Appendix D), suggesting that news media teach their consumers about these phrases, or perhaps have the effect of normalizing their usage. Of course, high rates of news consumption also connote high levels of political interest, which is also a significant predictor of outspokenness (and whose estimate becomes larger and more precise when news is excluded, as does the coefficient on social media usage, see Table 11), consistent with Barberá's (2015) interpretation of this parameter in the Tweetscores model of ideology. Ideological identity strength (which is the absolute value of the signed measure of ideological identity used in Figure 4) is also a significant predictor of outspokenness, but issue extremity (the absolute value of issue ideology used in Figure 4) is negative-signed, and not significantly different from zero,

implying that strong ideological identity, and *not* extreme issue ideology, provokes individuals to use catchphrases at all – which lends some nuance to the findings regarding identity and issue ideology reported above in relation to the α speech ideology trait. The coefficient on male gender is robustly null, which may be surprising in the context of Karpowitz and Mendelberg's 2014 studies of gender dynamics in deliberative settings, although it bears repeating that Wordsticks' β parameter captures not loquaciousness (as its Wordfish counterpart does) but the particular propensity to use politically-charged catchphrases when speaking about politics. Finally, in some specifications identifying as a person of color is significantly associated with outspokenness, but this is highly sensitive to the covariates included in the model.

In the rightmost panel of Figure 5, I plot the causal effects, in both the α and β dimensions, of the social context manipulations embedded in the WWYS question prompt in both studies. Compared to the common contrast of the close-friend context treatment that was included in both studies, I find that while the someone-you-just-met treatment (or "stranger" treatment, applied in Study 1, and plotted in pink) has no significant effect on participants' speech ideal point α, it induces a *downward* shift in outspokenness of greater magnitude than that associated with a 1-standard-deviation increase in news consumption or of political interest, which are the two strongest descriptive predictors of outspokenness. Substantively, this means that when speaking with strangers, individuals censor (or greatly reduce their usage of) all ideological catchphrases when speaking with strangers, as opposed to close friends.

Compared to the same close-friend reference point, I find that the "social media" treatment (applied in Study 2, and plotted in blue) induces both a *downward* and *rightward* shift in the speech space; that is, when going from their close friends to their online networks, social media posters censor their use of catchphrases generally, but the ones they do use connote a more right-wing speech ideal point than they express amongst their close friends. This may come as a surprise, considering that online spaces such as Twitter are often perceived to be unrepresentatively left-leaning, however I emphasize that this rightward shift observed amongst the "posters" does not take into account the fact that (as evidenced below) posters are generally more left-leaning than lurkers, nor does it take into account baseline differences in the ideological orientations of social media users *versus* non-users (who were not included in Study 2 at all).

It may also be surprising that people are less outspoken online than amongst their close friends, since many theorize that online platforms offer venues for people to to be more outspoken about politics than they can be amongst friends whose political views differ from their own. However, in a supplementary analysis (see Table 13, column 5), I interact the social media treatment with an indicator for whether the respondent perceives their online network to be more or less likeminded than their close friends, and recover a large and significant positive interaction effect. So, when posters perceive that their close-friend networks and online networks differ in likemindedness, they are significantly more outspoken in whichever setting they perceive to be more likeminded, consistent with longstanding theories of self-censorship (e.g. Noelle-Neumann, 1991). However, the majority of posters (78%) report that their friends and online networks are *both* likeminded, and display the same main effect wherein they are significantly less outspoken online than amongst close friends.

However, these main effects analyses do not address the question of ideological polarization in

online speech; to do this, I analyze the spreads of the distributions of speech ideology α in three exhaustive subsets of the Study 2 respondents, visualized in Figure 6: I plot the distribution of lurkers' close-friend α in gray, posters' close-friend α in black, and posters' online α in blue. That posters' speech ideologies are more polarized than lurkers' is visually evident from the broader spread of the black and blue distributions, relative to the gray, and is statistically verified by an F test for difference in variances ($p<0.0005$ for both comparisons). There is no evidence that posters adopt more polarized speech when posting their views online than when discussing them with close friends ($p=0.87$). This confirms the perception that online speech is unnaturally extreme, and supports the specific theory that this extremity is attributable to self-selection, and *not* ideological code-switching. Rather, as described above, the adjustments people make to their speech, when moving from offline to online contexts, tend to *censor* all political catchphrases.

Considering now the centers of the distributions, there is evidence that posters' close-friend speech ideology is significantly more leftist than that of lurkers ($p=0.021$). However, due to the rightward ideological code-switching induced by the social media treatment (discussed above), there is no significant difference in means between lurkers' close-friend speech ideology and posters' posted speech ideology ($p=0.88$). Posters' code-switching thus offsets the descriptive difference between lurkers' and posters' mean close-friend speech ideology, so that posters' online speech is not significantly to the left or right of lurkers' offline speech in general, even though it is significantly more polarized towards both the left and right extremes.

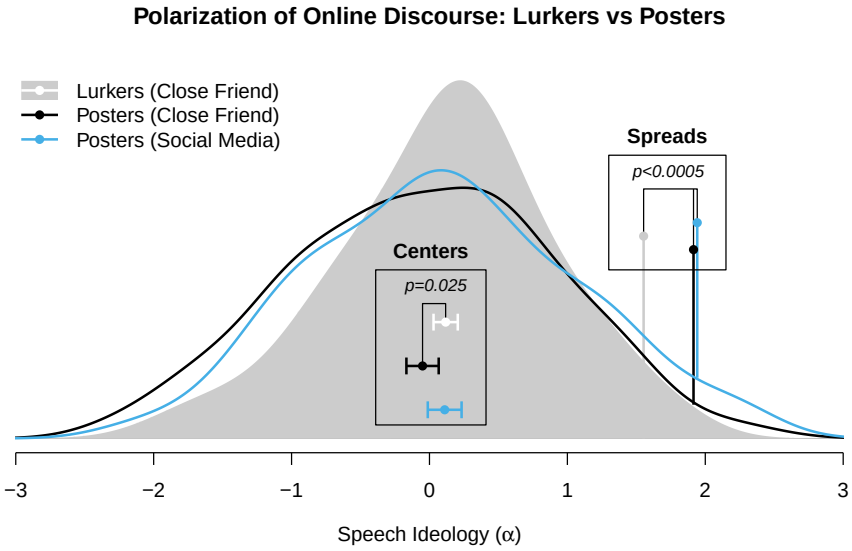

Figure 6: Density plots of respondent speech ideology α, in three exhaustive subsets of the Study 2 data: lurkers speaking with a close friend (gray), posters speaking with a close friend (black), and posters posting on their preferred online platform (blue).

7 Discussion

This paper has introduced a novel method for studying mass political speech, in which I adapt a model originally designed for scaling party manifestos, to estimate speech ideal points for members of the mass public based on their self-reported propensity to use certain political catchphrases. My method produces estimates of mass speech ideology that correlate strongly with other conventional measures of ideology, but also extend past findings about the relationship between ideological orientation and demographic variables like age, gender, and education, and lend insight into pressing contemporary questions about the causes of polarized political discourse on social media platforms.

It is perhaps surprising that this procedure works, considering that conventional text analysis methods typically employ much larger datasets than mine. The effectiveness of my method apparently reflects my participants' explicit knowledge of the political implications of the words and phrases I selected for this study, which makes their responses rich with information. I ask people to self-report speech behaviors, and their responses reflect a systematic ideological understanding of catchphrases. My "What Would You Say?" question may not be so different, after all, from the standard issue preference battery with which it so strongly correlates (*contra* concerns that catchphrases are policy-vapid signals of identity). But note how much more concise the "What Would You Say?" approach is: whereas the issue battery uses 136 words to pose 10 attitudinal questions (see Appendix K), the WWYS question poses twice as many questions using only 39 words (in the example shown in Figure 1). This parsimony perhaps reflects how these catchphrases function as shorthands, and the broader role of informal conversation in helping ordinary people make sense of complex political topics (c.f. Gamson, 1992).

At the same time, I do find that ideological social identity is a significant predictor of speech ideology, independent of concrete issue preferences. Of course, it is not necessarily a bad thing for people to use their speech to express overall identities, separate from positions on specific issues. My findings are consistent with Cramer Walsh's (2004) analysis of political conversation as a means to develop and refine one's political identity. Still, this may not completely allay concerns about the fundamental meaningfulness of catchphrases. Indeed, though issue ideology explains much more variance in speech ideology than does ideological identity, we might well wonder whether this issue battery is actually measuring the "true preferences" that we seek to distinguish from "cheap talk." After all, an issue battery is just more words, which the respondent claims to "agree" or "disagree" with. Viewed this way, the agree-disagree issue battery is just a more verbose way of asking, "What Would You Say?" in the guise of asking "What Would You Agree With?"

Scholars of public opinion have long criticized the conventional survey approach to attitude measurement (Sanders, 1999, see, e.g.), which unrealistically treats every issue as though it were up for public referendum by secret ballot. In real life, issues do not arrive as neat parcels at our doorsteps, discreetly subjecting themselves to our yea or nay within the privacy of our own minds; they arise subtly, even implicitly, often unexpectedly and sometimes awkwardly, but most of all *publicly*: when we express our opinions outside the artifice of a survey, we are almost always expressing ourselves *to somebody*. Where the standard issue battery erases this inherent sociality of public opinion (Blumer, 1948), the WWYS question brings it front-and-center, by specifying the

social context[14] of expression explicitly.

Furthermore, by rendering the social context of speech amenable to experimental manipulation, my method reveals causal effects on speech ideology (giving some support to theories of preference falsification, e.g. Kuran 1987), and to an even greater extent on outspokenness – a quantity estimated by my Wordsticks model, that remedies another long-lamented (Ginsberg, 1986) shortcoming of traditional public opinion measurement: the difficulty of characterizing respondents' intensity of feeling and commitment to advocate for their positions in everyday life.

I find that social context has large and significant effects on outspokenness, such that people are far more outspoken amongst close friends than when speaking with new acquaintances or posting on social media platforms. Moreover, I find that the effect of the social media context, relative to the close friend context, is significantly moderated by users' perceptions of the relative likemindedness of these environments: although the great majority perceive both as likeminded (which is to be expected, given humans' natural tendency to homophily), those who perceive one to be likeminded and the other not to be so, are significantly more outspoken in the likeminded setting. This offers new evidence in favor of Noelle-Neumann's (1991) theory that people use their "quasi-statistical sense" of others' opinions to guide their strategic self-censorship of unpopular views – which I am now able to elaborate in a richer linguistic framework of political expression. Noelle-Neumann originally dismissed the idea that people would adjust themselves to specific groups or social settings, so my research design and findings represent a departure from her original framework. However, it has long been theorized that homogeneous enclaves provide a space for individuals to express themselves more fully, and my findings appear to support this view.

These findings are all the more remarkable, considering that β was originally included in Slapin and Proksch's Wordfish model as a mere nuisance parameter, to account for variation in document length. By abjuring raw text altogether, my "What Would You Say?" measurement approach recovers substantive meaning in this parameter, and thus (perhaps ironically) derives richer insights into the nature and origins of political speech than would be gleaned from estimating the same type of model on data derived from natural language documents.

Of course, this study is not without limitations. First, in order to consider the results substantively meaningful, we must believe that respondents accurately report their propensity to say phrases. Although the accuracy of self-reports is commonly assumed in survey research, this does represent a distinct disadvantage for studying speech, relative to observational text analyses that measure attributes of individuals' actual speech (provided that documentation of their speech is available for analysis, of course). Before seeking to publish this paper, I intend to conduct a validation exercise in which I verify that the estimates of speech ideology derived from this method correlate significantly with estimates of the ideology of tweets obtained from the same individuals (using a method of tweet ideology measurement developed in a parallel project).

Second, some phrases' usage patterns are better-described by my Wordsticks model than others', which may raise questions about the dimensionality of the speech ideology construct – it may be valuable to estimate a version of Wordsticks that supposes two dimensions of speech ideology,

[14] One could specify a social context in a traditional agree-disagree issue battery, but doing so would not evoke a realistic social situation.

rather than one. This might recover the familiar social and economic dimensions often thought to characterize American political ideology, whereas the single dimension returned by the current approach seems primarily social in nature. Indeed, although the phrases "estate tax" and "death tax," were included in my studies specifically because they so notably distinguished Democratic from Republican legislators in Gentzkow and Shapiro's (2010) study, they receive some of the weakest γ discrimination parameters of all the phrases I scaled. While these terms might prove more discriminative in a two-dimensional ideology space, it is also possible that the mass public's speech patterns are simply different from elites', and that these weak discrimination parameters are valid (future studies might investigate differences in the ideological content of elite and mass speech). And though most phrases satisfy the monotonicity assumption encoded in Wordsticks' IRT specification (see Appendix H), some of the exceptions are substantively sensible: "woke," for example, is a term once exclusively used sincerely by the Left, and now increasingly used derogatorily by the Right; hence its usage is perhaps truly non-monotonic on speech ideology. Specifying a version of Wordsticks that allows for multi-peaked usage curves is beyond the scope of this paper, but would be a valuable extension.

Third, the choice of phrases to include in the WWYS question is inherently subjective, and the results may depend on the choice of phrases. Although I take some reassurance from the fact that the results reported in this paper are derived from a data collection approach in which respondents responded to a subset of 20 phrases randomly sampled from a superset of 46 phrases (and so cannot be dependent on the respondent seeing a specific set of 20 phrases), it remains plausible that a study using a different total set of phrases would return substantively different findings. However, it is easy to test this proposition: if the reader doubts these results' validity due to my selection or omission of certain phrases, it is quite straightforward to conduct a replication study using alternative phrases and observe the consequences. Similarly, while the present study is deliberately US-centric in its choice of phrases, I eagerly invite scholars to apply my design to study catchphrases in other settings and languages. So, while the subjectivity of phrase selection may impose substantive limits on any particular study using this method, it also creates many opportunities for future applications of the method in general: scholars can apply their substantive knowledge of politics to study any forms of political speech that interest them, on any topic, in any language in which the researcher is politically-fluent.

Having introduced this new approach and demonstrated its capabilities, I now conclude with the hope it will be of use to text analysis practitioners, and to other scholars who seek to understand how members of the mass public use systems of shared linguistic meaning to express their ideals and identities.

Bibliography

AMA. 2021. "Advancing Health Equity: A Guide to Language, Narrative and Concepts.".
URL: *https://www.ama-assn.org/about/ama-center-health-equity/advancing-health-equity-guide-language-narrative-and-concepts-0*

Arceneaux, Kevin and Rory Truex. 2021. Donald Trump and the Lie. preprint PsyArXiv.
URL: *https://osf.io/e89ym*

Barbera, Pablo. 2016. "Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data." p. 37.

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost and Joshua A Tucker. 2019. "Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data." *American Political Science Review* 113(4):883–901.

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data." *Political Analysis* 23(1):76–91.

Bartholomew, James. 2015. "Easy Virtue." *The Spectator* .
URL: *https://www.spectator.co.uk/article/easy-virtue*

Bicchieri, Cristina. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.
URL: *http://ebooks.cambridge.org/ref/id/CBO9780511616037*

Blumer, Herbert. 1948. "Public Opinion and Public Opinion Polling." *American Sociological Review* 13(5):542.
URL: *http://www.jstor.org/stable/2087146?origin=crossref*

Bor, Alexander and Michael Bang Petersen. 2021. "The Psychology of Online Political Hostility: A Comprehensive, Cross-National Test of the Mismatch Hypothesis." *American Political Science Review* pp. 1–18.
URL: *https://www.cambridge.org/core/product/identifier/S0003055421000885/type/journal$_a$rticle*

Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker and Jay J. Van Bavel. 2017. "Emotion shapes the diffusion of moralized content in social networks." *Proceedings of the National Academy of Sciences* 114(28):7313–7318.
URL: *http://www.pnas.org/lookup/doi/10.1073/pnas.1618923114*

Brady, William J., Killian McLoughlin, Tuan N. Doan and Molly J. Crockett. 2021. "How social learning amplifies moral outrage expression in online social networks." *Science Advances* 7(33):eabe5641.
URL: *https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.abe5641*

Butler, Judith. 1999. *Gender trouble: Feminism and the subversion of identity*. Routledge.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2):355–370.
URL: *https://www.cambridge.org/core/product/identifier/S0003055404001194/type/journal$_a$rticle*

Coaston, Jane. 2017. "'Virtue Signaling' Isn't the Problem. Not Believing One Another Is." *The New York Times* .
URL: *https://www.nytimes.com/2017/08/08/magazine/virtue-signaling-isnt-the-problem-not-believing-one-another-is.html*

Cramer Walsh, Katherine. 2004. *Talking about politics: Informal groups and social identity in american life.* Chicago: University of Chicago Press.

Cuthbert, Lane and Alexander Theodoridis. 2022. "Analysis | Do Republicans really believe Trump won the 2020 election? Our research suggests that they do." *Washington Post* .
URL: *https://www.washingtonpost.com/politics/2022/01/07/republicans-big-lie-trump/*

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv:1810.04805 [cs]* . arXiv: 1810.04805.
URL: *http://arxiv.org/abs/1810.04805*

DiAngelo, Robin J. 2021. *Nice racism: how progressive white people perpetuate racial harm.* OCLC: 1262180071.

Errington, Joseph. 1999. "Ideology." *Journal of Linguistic Anthropology* 9(1/2):115–117. Publisher: [American Anthropological Association, Wiley].
URL: *https://www.jstor.org/stable/43102441*

Gamson, William A. 1992. *Talking politics.* Cambridge [England] ; New York, NY, USA: Cambridge University Press.

Gentzkow, Matthew and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." *Econometrica* 78(1):35–71.

Gilligan, Carol. 1993. *In a different voice: psychological theory and women's development.* Cambridge, Mass: Harvard University Press.

Ginsberg, Benjamin. 1986. *The Captive Public: How Mass Opinion Promotes State Power.* New York: Basic Books.

Goffman, Erving. 1956. *The Presentation of Self in Everyday Life.* Number 2 *in* "University of Edinburgh Social Sciences Research Centre Monographs" Edinburgh: University of Edinburgh Press.

Goldberg, Michelle. 2019. "Twitter Isn't Real Life (if You're a Democrat)." *The New York Times* .
URL: *https://www.nytimes.com/2019/05/13/opinion/joe-biden-twitter-2020.html*

Goplerud, Max. 2019. "A Multinomial Framework for Ideal Point Estimation." *Political Analysis* 27(1):69–89.
URL: *https://www.cambridge.org/core/product/identifier/S1047198718000311/type/journal$_a$rticle*

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.

Guess, Andrew, Brendan Nyhan, Jin Woo Kim and Jason Reifler. 2019. "The Distorting Prism of Social Media: How Online Comments Amplify Toxicity.". tex.ids: Guess2019a.

Hanson, Victoria Davis. 2020. "A Guide to Wokespeak.".
URL: *https://www.nationalreview.com/2020/12/a-guide-to-wokespeak/*

Harmon, Amy. 2021. "BIPOC or POC? Equity or Equality? The Debate Over Language on the Left." *The New York Times* .
URL: *https://www.nytimes.com/2021/11/01/us/terminology-language-politics.html*

Huddy, Leonie, Lilliana Mason and Lene Aarøe. 2015. "Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity." *American Political Science Review* 109(1):1–17.
URL: *https://www.cambridge.org/core/product/identifier/S0003055414000604/type/journal$_a$rticle*

Jaidka, Kokil. 2021. "Discussion Quality Measurement on Social Media: Developing and Validating Dictionaries Based on an Open Vocabulary Approach." *SSRN Electronic Journal* .
URL: *https://www.ssrn.com/abstract=3870554*

Kaba, Mariame. 2020. "Yes, We Mean Literally Abolish the Police." *The New York Times* .
URL: *https://www.nytimes.com/2020/06/12/opinion/sunday/floyd-abolish-defund-police.html*

Karpowitz, Christopher and Tali Mendelberg. 2014. *The Silent Sex: Gender, Deliberation, and Institutions*. Princeton: Princeton University Press.

Katz, Elihu and Paul F Lazarsfeld. 1955. *Personal influence: the part played by people in the flow of mass communications.* New York, NY, US: Free Press. Pages: xx, 400 Publication Title: Personal influence: the part played by people in the flow of mass communications.

King, Gary, Benjamin Schneer and Ariel White. 2017. "How the news media activate public expression and influence national agendas." *Science* 358(6364):776–780.
URL: *https://science.sciencemag.org/content/358/6364/776*

Klašnja, Marko, Pablo Barberá, Nicholas Beauchamp, Jonathan Nagler and Joshua A. Tucker. 2015. *Measuring public opinion with social media data.* Issue: July 2019 Pages: 582 Publication Title: The Oxford Handbook of Polling and Polling Methods.

Kopan, Tal. 2018. "Justice Department: Use 'illegal aliens,' not 'undocumented' | CNN Politics.".
URL: *https://www.cnn.com/2018/07/24/politics/justice-department-illegal-aliens-undocumented/index.html*

Kuran, Timur. 1987. "Preference Falsification , Policy Continuity and Collective Conservatism." *The Economic Journal* 97:642–665.

Kuran, Timur. 1995. *Private Truths, Public Lies: The Social Consequences of Preference Falsification. Cambridge, MA: Harvard University Press.* Harvard University Press.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(02).
URL: *http://www.journals.cambridge.org/abstract$_S$0003055403000698*

Lowe, Will. 2001. "Towards a Theory of Semantic Space." *Proceedings of the Annual Meeting of the Cognitive Science Society* 23(23):7.

Lowrey, Annie. 2020. "What "Defund the Police" Actually Means.". Section: Ideas.
URL: *https://www.theatlantic.com/ideas/archive/2020/06/defund-police/612682/*

Mason, Lilliana. 2018. *Uncivil Agreement: How Politics Became Our Identity.* Chicago: University of Chicago Press.

McConnell-Ginet, Sally. 2020. *Words Matter: Meaning and Power.* 1 ed. Cambridge University Press.
URL: *https://www.cambridge.org/core/product/identifier/9781108641302/type/book*

McWhorter, John. 2021. "Opinion | Honestly, You Should Keep Using These Verboten Terms." *The New York Times* .
URL: *https://www.nytimes.com/2021/10/12/opinion/language-words-woke.html*

Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16(4):372–403.
URL: *https://www.cambridge.org/core/product/identifier/S1047198700002291/type/journal$_a$rticle*

Nguyen, Thu T, Nikki Adams, Dina Huang, M Maria Glymour, Amani M Allen and Quynh C Nguyen. 2020. "The Association Between State-Level Racial Attitudes Assessed From Twitter Data and Adverse Birth Outcomes: Observational Study." *JMIR Public Health and Surveillance* 6(3):e17103.
URL: *https://publichealth.jmir.org/2020/3/e17103*

Noel, Hans. 2012. "The Coalition Merchants: The Ideological Roots of the Civil Rights Realignment." *The Journal of Politics* 74(1):156–173.
URL: *https://www.journals.uchicago.edu/doi/10.1017/S0022381611001186*

Noelle-Neumann, Elisabeth. 1991. The Theory of Public Opinion: The Concept of the Spiral of Silence. In *Communication Yearbook.* International Communication Association pp. 256–287.

Noelle-Neumann, Elisabeth. 1974. "The Spiral of Silence A Theory of Public Opinion." *Journal of Communication* 24(2):43–51.

Pagel, Mark, Mark Beaumont, Andrew Meade, Annemarie Verkerk and Andreea Calude. 2019. "Dominant words rise to the top by positive frequency-dependent selection." *Proceedings of the National Academy of Sciences* 116(15):7397–7402. Publisher: National Academy of Sciences Section: Biological Sciences.
URL: *https://www.pnas.org/content/116/15/7397*

Poole, Keith T. and Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29(2):357.
URL: *https://www.jstor.org/stable/2111172?origin=crossref*

Rinberg, Toly, Maya Anjur-Dietrich, Marcy Beck, Andrew Bergman, Justin Derry, Lindsey Dillon, Gretchen Gehrke, Rebecca Lave, Chris Sellers, Nick Shapiro, Anastasia Aizman, Dan Allan, Madelaine Britt, Raymond Cha, Janak Chadha, Morgan Currie, Sara Johns, Abby Klionsky, Stephanie Knutson, Katherine Kulik, Aaron Lemelin, Kevin Nguyen, Eric Nost, Kendra Ouellette, Lindsay Poirier, Sara Rubinow, Justin Schell, Lizz Ultee, Julia Upfal, Tyler Wedrosky and Jacob Wylie. 2018. Changing the Digital Climate. Technical report Environmental Data & Governance Initiative.
URL: *https://envirodatagov.org/wp-content/uploads/2018/01/Part-3-Changing-the-Digital-Climate.pdf*

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: STRUCTURAL TOPIC MODELS FOR SURVEY RESPONSES." *American Journal of Political Science* 58(4):1064–1082.
URL: *http://doi.wiley.com/10.1111/ajps.12103*

Sanders, Lynn M. 1999. "Democratic Politics and Survey Research." *Philosophy of the Social Sciences* 29(2):248–280.
URL: *http://journals.sagepub.com/doi/10.1177/004839319902900205*

Schulz, William Small, Andrew M. Guess, Pablo Barberá, Simon Munzert, JungHwan Yang, Adam Hughes, Emma Remy, Sono Shah and Aaron Smith. 2020. "(Mis)representing Ideology on Twitter: How Social Influence Shapes Online Political Expression." Working paper presented at APSA 2020.

Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman and Lyle H. Ungar. 2013. "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach." *PLoS ONE* 8(9):e73791.
URL: *https://dx.plos.org/10.1371/journal.pone.0073791*

Settle, Jamie E. 2018. *Frenemies: How Social Media Polarizes America*. New York: Cambridge University Press.

Shear, Michael D. 2021. "The Words That Are In and Out With the Biden Administration." *The New York Times* .

URL: *https://www.nytimes.com/2021/02/24/us/politics/language-government-biden-trump.html*

Silverstein, Michael. 2003. "Indexical order and the dialectics of sociolinguistic life." *Language & Communication* 23(3-4):193–229.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0271530903000132*

Simonovits, Gabor. 2017. "Centrist by Comparison: Extremism and the Expansion of the Political Spectrum." *Political Behavior* 39(1):157–175.
URL: *http://link.springer.com/10.1007/s11109-016-9351-y*

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/j.1540-5907.2008.00338.x*

Sun, Lena H. and Juliet Eilperin. 2017. "CDC gets list of forbidden words: Fetus, transgender, diversity." *Washington Post* .
URL: *https://www.washingtonpost.com/national/health-science/cdc-gets-list-of-forbidden-words-fetus-transgender-diversity/2017/12/15/f503837a-e1cf-11e7-89e8-edec16379010*$_s$*tory.html*

Sunstein, Cass R. 2017. *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press. tex.ids= Sunstein2017RepublicDivideda.

Traugott, E.C. 2006. Semantic Change: Bleaching, Strengthening, Narrowing, Extension. In *Encyclopedia of Language & Linguistics*. Elsevier pp. 124–131.
URL: *https://linkinghub.elsevier.com/retrieve/pii/B0080448542011056*

Tufekci, Zeynep. 2017. *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

Turner, John C. 1991. *Social influence*. Bristol, PA: Open University Press.

Wardhaugh, Ronald and Janet M. Fuller. 2015. Chapter 3: Defining Groups. In *An introduction to sociolinguistics*. Seventh edition. ed. Blackwell textbooks in linguistics. West Sussex, England: John Wiley & Sons pp. 62–79.

Warzel, Charlie. 2020. "Twitter Is Real Life." *The New York Times* .
URL: *https://www.nytimes.com/2020/02/19/opinion/twitter-debates-real-life.html*

Wojcik, Stefan and Adam Hughes. 2019. Sizing Up Twitter Users. Technical report Pew Research Center.
URL: *https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/04/twitter*$_o$*pinions*$_{41}$8$_f$*inal*$_c$*lean.pdf*

Yglesias, Matthew. 2020. "Defund police is a bad idea, not a bad slogan.".
URL: *https://www.slowboring.com/p/defund-police-is-a-bad-idea-not-a*

Yglesias, Matthew. 2021. “The AMA’s "Advancing Health Equity" plan leaves out everything that matters.”.
URL: *https://www.slowboring.com/p/the-amas-advancing-health-equity*

Zito, Salena. 2016. “Taking Trump Seriously, Not Literally.”. Section: Politics.
URL: *https://www.theatlantic.com/politics/archive/2016/09/trump-makes-his-case-in-pittsburgh/501335/*

A Wordsticks Model

In order to estimate a respondent latent trait called "speech ideology" from responses to the What Would You Say question, I specify a spatial choice model, encoding several key assumptions:

- Respondents i can be placed along a left-right spectrum of speech ideology, modeled as a normally-distributed latent trait α_i.
- Probability of saying each phrase j is either strictly increasing or strictly decreasing in speech ideology, such that for each phrase j, two cutpoints c_j^2 and c_j^3 can exhaustively partition the speech ideology line into a region in which "I would say this" is the most probable response, a region in which "I would not say this" is the most probable response, and an intermediate region in which "I might say this" is the most probable response.
- The direction and rate of change in probability of saying each phrase j, as a function of respondent speech ideology α_i, varies across phrases, and can be summarized with a scalar γ_j, which multiplies the distance between the respondent's speech ideology and cutpoint c_j^k. The sign of γ_j denotes whether phrase j is left- or right-slanted, by determining the direction of the speech ideology space in which phrase usage is increasing. The magnitude of γ_j corresponds to the extent to which usage of phrase j correlates with speech ideology.
- Respondents also vary in their baseline propensity to say any phrase. This can be interpreted as respondents having different interpretations of the response scale, or having different levels of desire to engage in this kind of political speech, independent of ideology. It is modeled as an additive intercept β_i.

Formally, for respondents indexed by i and phrases indexed by j,

$$\Pr(y_{ij} \geq k) \ = \frac{\exp(\mu_{ij}^k)}{1+\exp(\mu_{ij}^k)}; \qquad \mu_{ij}^k = (\alpha_i - c_j^k) \times \gamma_j + \beta_i$$

Where

y_{ij} the observed response to phrase j reported by respondent i
$k \in \{2, 3\}$ the possible response categories (excluding 1, the lowest category)
α_i the latent speech ideology of respondent i
c_j^k the ideology cutpoint between response $k-1$ and k for phrase j
γ_j the ideological slant of phrase j
β_i the tendency of respondent i to indicate greater propensity to say phrases, net of ideology

Note also the following, which shows that the spatial choice model can be rearranged to an equivalent specification that is both more convenient to estimate, and that also makes clear how the inclusion of cutpoints c_j renders a phrase-level intercept, such as Slapin and Proksch (2008) included in Wordfish, redundant.

$$
\begin{aligned}
\mu_{ij}^{k} &= (\alpha_i - c_j^k) \times \gamma_j + \beta_i \\
&= \alpha_i \times \gamma_j - c_j^k \times \gamma_j + \beta_i \\
&= \alpha_i \times \gamma_j + \beta_i - c_j^k \times \gamma_j \\
&= \alpha_i \times \gamma_j + \beta_i - c_j^{k\prime}, \quad \text{where } c_j^{k\prime} = c_j^k \times \gamma_j
\end{aligned}
$$

The latter formulation is implemented in the STAN modeling language in Appendix B. This formulation also makes clear that Wordsticks is a straightforward ordinal extension of Wordfish, as originally specified by Slapin and Proksch (2008):

$$
\Pr(y_{ij} = k) \ = \frac{(\lambda_{ij})^k \times \exp(-\lambda_{ij})}{k!}; \qquad \lambda = \exp(\mu_{ij}^{'}); \qquad \mu_{ij}^{'} = \alpha_i \times \gamma_j + \beta_i + \theta_j
$$

Although Wordfish uses a Poisson distribution to model the observed data as a count, the Wordfish running variable $\mu_{ij}^{'}$ is nearly identical to Wordsticks' running variable μ_{ij}^{k}, except that Wordfish estimates a word-level intercept parameter θ_j that takes a single value for each word, whereas Wordsticks estimates two phrase-level parameters c_j^2 and c_j^3 that correspond to the cutpoints between the three ordinal response categories (but which are otherwise functionally the same as the Wordfish θ_j).

B STAN Code for Wordsticks Model

```
data {
  int<lower=1> N;                  //number of data points
  int<lower=1> J;                  //number of subjects
  int<lower=1> K;                  //number of items
  int<lower=1,upper=3> say[N];     //outcome: "I would say this"/"I might say this"/"I would not say this"
  int<lower=1, upper=J> subj[N];   //subject id
  int<lower=1, upper=K> item[N];   //item id
}

parameters {
  vector[J] alpha_raw;         // respondent speech ideology (raw)
  vector[J] beta;              // respondent "outspokenness"
  vector[K] gamma;             // item slants (direction and strength of association with speech ideology)
  vector[K] c_location;        // cutpoint location parameters
  vector<lower=0>[K] c_scale;  // cutpoint scale parameters
  real mu_gamma;               // slant mean
  real<lower=0.1> sigma_gamma; // slant spread
  real<lower=0.1> sigma_beta;  // outspokenness spread
}

transformed parameters { // hard-standardize alphas to aid model identification
  vector[J] alpha;
  alpha = (alpha_raw - mean(alpha_raw)) ./ sd(alpha_raw);
}

model {
  vector[2] c;
  //priors
  gamma[6] ~ exponential(.1);            // enforce America first right-slanted
  gamma ~ normal(mu_gamma,sigma_gamma);  // normal prior on slants
  alpha_raw ~ normal(0,1);               // standard normal prior on speech ideology
  beta ~ normal(0,sigma_beta);           // normal prior (mean zero) on phrase slants
  c = [ -1 , 1 ]';                       // base cutpoints (transformed by location and scale params)

  for (i in 1:N){
    say[i] ~ ordered_logistic(
    gamma[item[i]] * alpha[subj[i]] + beta[subj[i]],   // utility function
    c*c_scale[item[i]] + c_location[item[i]]);           // response cutpoints
  }
}

generated quantities {  // transform cutpoints into ideology space for plotting
  vector[K] c_transformed_1;
  vector[K] c_transformed_2;
  c_transformed_1 = (-1 * c_scale + c_location) ./ gamma;
  c_transformed_2 = (1 * c_scale + c_location) ./ gamma;
}
```

C Regression Tables & Detailed Discussion

C.1 Alpha Descriptive Analyses

Table 10 presents descriptive analyses to identify predictors of α speech ideology: The first column represents study 1, which did not include an issue scale, but included an ideological identity scale. I find that ideological identity only predicts speech ideology when 6-point self-described ideology is excluded from the model (column 2), which may be a consequence of the collinearity of these two variables. Since the sign of ideological identity is constructed directly from responses to the 6-point item, I subsequently avoid including both measures in the same model, to aid substantive interpretation of the coefficients. Comparing this to study 2 (column 3), which did include an issue ideology scale, I find issue ideology largely displaces 6-point ideology's explanatory power, but that ideological identity strength remains a strongly significant predictor of speech ideology.

Turning to consider demographic correlates of speech ideology, in study 1 (columns 1-2), college education is associated with more left-leaning speech, male gender (as compared to female and non-binary gender) is associated with more right-leaning speech, and older individuals tend to speak more conservatively. Study 2 (columns 3-4) replicated these findings, with the exception that college education, receives a null coefficient when issue ideology is present in the model (column 3). When pooling the data across both studies (columns 5-6), college education remains a significant predictor of speech ideology. Note that column 5 presents a pooled model in which only the conventional likert-scale measure of ideology is included, excluding both the identity and issue measures, while column 6 presents a fully-saturated model (note that issue ideology cannot be included in a pooled model since it was not measured in Study 1).

Table 1: Alpha Descriptive Analyses

	Pilot 1 (Prolific)	Pilot 1 w/o SDI	Pilot 2 (MTurk)	Pilot 2 w/o Issues	Pooled w/o Identity	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
age_dec	0.098***	0.105***	0.070***	0.093***	0.091***	0.093***
	(0.024)	(0.024)	(0.015)	(0.014)	(0.014)	(0.014)
sdi_num6	0.607***			0.457***	0.630***	0.457***
	(0.125)			(0.077)	(0.052)	(0.077)
pid_num6	0.321***	0.452***	0.154***	0.329***	0.374***	0.329***
	(0.078)	(0.075)	(0.053)	(0.049)	(0.047)	(0.049)
pol_int	−0.068	−0.137	−0.156**	−0.136**	−0.118*	−0.136**
	(0.109)	(0.111)	(0.074)	(0.067)	(0.067)	(0.067)
issue_scale			0.493***			
			(0.029)			
signed_identity	0.070	0.359***	0.164***	0.148***		0.148***
	(0.079)	(0.054)	(0.038)	(0.049)		(0.049)
college	−0.190**	−0.193**	0.009	−0.108**	−0.114**	−0.108**
	(0.075)	(0.077)	(0.051)	(0.046)	(0.046)	(0.046)
POC	0.005	0.020	−0.051	−0.027	−0.027	−0.027
	(0.072)	(0.073)	(0.050)	(0.045)	(0.045)	(0.045)
male	0.185***	0.182***	0.111***	0.180***	0.180***	0.180***
	(0.056)	(0.057)	(0.036)	(0.034)	(0.034)	(0.034)
media_scale	−0.053*	−0.037	0.030	−0.0001	−0.001	−0.0001
	(0.030)	(0.030)	(0.020)	(0.018)	(0.018)	(0.018)
sm_scale	−0.015	−0.017	0.016	−0.005	−0.008	−0.005
	(0.027)	(0.028)	(0.023)	(0.019)	(0.019)	(0.019)
Constant	−0.272**	−0.263**	−0.182**	−0.229***	−0.224***	−0.229***
	(0.128)	(0.130)	(0.088)	(0.079)	(0.079)	(0.079)
Observations	502	502	797	1,299	1,299	1,299
R^2	0.581	0.561	0.697	0.584	0.581	0.584
Adjusted R^2	0.572	0.553	0.693	0.580	0.578	0.580

Note: *p<0.1; **p<0.05; ***p<0.01

C.2 Beta Descriptive Analyses

Table 11 presents results from regression analyses with respondent β as the dependent variable. Model 1 focuses on basic political and demographic variables, and finds that β is significantly associated with both political interest and ideological identity strength, consistent with the interpretation of β as a kind of political expressiveness orthogonal to left-right ideology. Model 2 finds that the absolute value of issue ideology has a significant *negative* association with the absolute value (or extremity) of issue ideology, which is perhaps surprising: we might expect people who hold more extreme political attitudes to be more likely to use politically-charged language. To investigate this further, model 3 drops political interest and identity strength, which results in a null coefficient on issue extremity – I interpret this as an extension of the findings reported above regarding α and ideological identity: using this kind of language is a way of expressing an affective investment in politics, distinct from concrete issue attitudes. After accounting for this affective attachment to politics, it seems that the ideological coherence of one's actual policy views is associated with more careful or guarded speech habits.

Models 4, 5, and 6 explore the relationship between outspokenness and media habits: model 4 introduces an index of news consumption, which is positively associated with outspokenness (Table 5, column 1, presents results from a disaggregated model, showing coefficients on individual news media types). Model 5 introduces an index of social media usage, which takes a smaller but still significant positive coefficient (Table 5, column 2, presents the disaggregated version). Model 6 includes both media scales, and both retain significant positive coefficients. These findings suggest that engaging with news media and social media are associated with outspokenness, possibly because these media inform citizens about the political connotations of these phrases, or normalize their usage by featuring individuals who speak in these ways. Model 7 presents a fully saturated specification.

Table 2: Beta Descriptive Analyses

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
age_dec	0.020	0.042	0.092***	0.019	0.047*	0.036	0.053
	(0.025)	(0.032)	(0.033)	(0.024)	(0.027)	(0.026)	(0.034)
sdi_num6	0.061	−0.019	0.001	−0.0004	0.076	0.010	−0.036
	(0.099)	(0.130)	(0.133)	(0.096)	(0.099)	(0.096)	(0.126)
pid_num6	−0.109	−0.112	−0.170	−0.066	−0.108	−0.066	−0.101
	(0.090)	(0.118)	(0.121)	(0.088)	(0.090)	(0.087)	(0.115)
pol_int	0.641***	0.865***		0.229*	0.584***	0.202	0.390**
	(0.127)	(0.167)		(0.131)	(0.128)	(0.132)	(0.177)
identity_strength	0.181***	0.154***		0.153***	0.178***	0.151***	0.110**
	(0.035)	(0.046)		(0.034)	(0.035)	(0.034)	(0.045)
college	−0.036	−0.106	−0.054	−0.091	−0.030	−0.086	−0.118
	(0.089)	(0.117)	(0.120)	(0.086)	(0.089)	(0.086)	(0.114)
POC	0.194**	0.142	0.079	0.138*	0.178**	0.129	0.100
	(0.086)	(0.114)	(0.117)	(0.084)	(0.086)	(0.084)	(0.111)
male	−0.009	0.054	0.081	−0.009	−0.005	−0.007	0.046
	(0.066)	(0.085)	(0.085)	(0.064)	(0.065)	(0.064)	(0.082)
abs(issue_scale)		−0.182**	0.039				−0.082
		(0.083)	(0.079)				(0.082)
media_scale				0.311***		0.304***	0.299***
				(0.034)		(0.034)	(0.045)
sm_scale					0.099***	0.063*	0.047
					(0.036)	(0.035)	(0.052)
Constant	−0.478***	−0.554***	−0.424**	−0.157	−0.556***	−0.213	−0.371*
	(0.146)	(0.201)	(0.195)	(0.146)	(0.148)	(0.149)	(0.203)
Observations	1,299	797	797	1,299	1,299	1,299	797
R^2	0.070	0.077	0.018	0.127	0.076	0.129	0.129
Adjusted R^2	0.065	0.067	0.009	0.121	0.069	0.123	0.117

Note: *p<0.1; **p<0.05; ***p<0.01

C.3 Alpha Treatment Effects

Table 12 presents regression analyses to estimate main effects of the treatments on the α speech ideology parameter. Columns 1 and 2 show no significant main effect of the "someone you just met" (or to be concise, "stranger") treatment on α ideology, indicating that people do not have a general tendency to adjust the ideology of their speech when talking to strangers as opposed to close friends. Columns 3 and 4 show a significant positive effect on the social media context treatment (or "smct"), indicating that individuals who express political beliefs online tend to adopt somewhat more conservative speech patterns online, compared to how they would speak with a close friend. This would seem to contradict the common perception that online speech is unrepresentatively left-leaning, however it should be noted that the coefficient on the social media context treatment reflects a code-switching effect, net of issue ideology, ideological identity, and the other factors discussed above. In fact, when we consider model 3, we can see descriptively that there is a baseline difference in the speech ideology of individuals who express their political views online in the first place: posters have more liberal speech ideology, and the magnitude of this coefficient is slightly larger than that of the code-switching effect, but in the opposite direction. This indicates that the people who speak up about politics online tend to have more liberal speech patterns when talking to close friends (due to some combination of left-leaning attitudes and identities), but their rightward code-switching largely erases this liberalism when moving to the online context.

Table 3: Alpha Treatment Effects

	(1)	(2)	(3)	(4)
age_dec		0.097***		0.071***
		(0.024)		(0.015)
sdi_num6		0.607***		−0.035
		(0.125)		(0.088)
pid_num6		0.322***		0.155***
		(0.078)		(0.055)
pol_int		−0.068		−0.160**
		(0.109)		(0.074)
issue_scale				0.496***
				(0.030)
signed_identity		0.069		0.179***
		(0.080)		(0.054)
college		−0.190**		0.002
		(0.075)		(0.051)
POC		0.006		−0.057
		(0.072)		(0.050)
male		0.185***		0.110***
		(0.056)		(0.036)
media_scale		−0.053*		0.033
		(0.030)		(0.020)
sm_scale		−0.016		0.017
		(0.027)		(0.023)
stranger	−0.043	−0.014		
	(0.082)	(0.054)		
expressor			−0.167**	−0.062
			(0.076)	(0.045)
smct_X_expressor			0.160**	0.119***
			(0.081)	(0.045)

C.4 Beta Treatment Effects

Table 13 presents regression analyses to estimate main effects of the treatments on the β outspokenness intercept parameter, revealing robust and substantively large negative effects of both the "stranger" treatment (models 1 and 2) and social media context treatment (models 3 and 4), relative to the "close friend" treatment. As can be seen in model 2, the stranger treatment takes a coefficient with magnitude roughly twice as great as any other coefficient in the model, and it is noteworthy that this model explains nearly twice as much of the variance in β as the fully-saturated models presented in Table 11. So, people appear to self-censor when talking to new acquaintances, compared to how they would speak with close friends. Further studies should probe the reasons for this effect, which may be related to concerns about self-presentation when meeting new people.

The social media context treatment takes a strongly significant negative coefficient, with a magnitude roughly half as great. Somewhat similar to the findings regarding treatment effects on the α parameter, posters are found to be generally more outspoken (and as we would expect, this effect seems to be explained by differences in political attitudes and identities), but the countervailing effect of the social media context treatment almost entirely erases this difference. Thus, people who speak up about politics online are generally more outspoken amongst friends (compared to lurkers speaking with friends), but exhibit profound self-silencing in their online speech. Note, however, the model presented in column 5, which subsets the Study 2 sample to include only the posters, and interacts the social media context treatment with a signed indicator variable that encodes whether the respondent perceives their online networks to be likeminded but not their close friends (1), perceives both groups to be likeminded (0), or perceives their friends to be likeminded but not their online networks (-1). This interaction between the social media context treatment, and the indicator representing relative likemindedness of one's social media environment *versus* one's friend group, takes a large and significantly positive coefficient. Substantively, this suggests (consistent with prevalent theories) that people who hold views different from their friends may be using social media platforms as alternative spaces, to cultivate the connections with likeminded people that they lack in their local networks, so that they can be more outspoken. Further studies should probe this further, as well as investigating other mechanisms behind this effect, which may also be moderated by personality traits such as self-monitoring.

It is also noteworthy that this self-censorship effect is greater in the stranger treatment than in the social media treatment, since (depending on the platform) online speech may be received by many people who are in fact strangers to the user who is speaking. Thus the more relevant comparison may be between face-to-face versus online speech to an audience of strangers, but this comparison is not afforded by the design of the two studies conducted so far.

Table 4: Beta Treatment Effects

	(1)	(2)	(3)	(4)	(5)
age_dec		0.021		0.051	0.036
		(0.041)		(0.034)	(0.041)
sdi_num6		0.029		−0.047	−0.089
		(0.144)		(0.126)	(0.149)
pid_num6		0.057		−0.091	−0.071
		(0.131)		(0.115)	(0.134)
pol_int		−0.021		0.365**	0.440**
		(0.203)		(0.171)	(0.216)
identity_strength		0.252***		0.110**	0.187***
		(0.053)		(0.044)	(0.054)
college		−0.029		−0.111	−0.089
		(0.130)		(0.114)	(0.137)
POC		0.160		0.119	0.271*
		(0.124)		(0.111)	(0.138)
male		−0.084		0.042	−0.039
		(0.098)		(0.082)	(0.099)
media_scale		0.286***		0.309***	0.261***
		(0.052)		(0.045)	(0.052)
sm_scale		0.086*		0.055	0.046
		(0.047)		(0.052)	(0.064)
stranger	−0.561***	−0.525***			
	(0.100)	(0.093)			
expressor			0.308***	0.019	
			(0.101)	(0.100)	
smct_X_expressor			−0.280***	−0.248**	−0.273***
			(0.106)	(0.100)	(0.097)
sm_friend_likemindedness_diff					−0.110
					(0.149)
smct_X_expressor:sm_friend_likemindedness_diff					0.449**
					(0.207)
Constant	0.272***	0.291	−0.092	−0.362*	−0.328
	(0.070)	(0.229)	(0.068)	(0.205)	(0.252)
Observations	503	502	798	797	498
R^2	0.059	0.210	0.014	0.136	0.172
Adjusted R^2	0.057	0.192	0.011	0.123	0.149

D Disaggregated Media Models

Table 5: Beta Disaggregated Media Effects

	(1)	(2)
tv	0.023	
	(0.019)	
newspapers	0.058***	
	(0.022)	
radio	0.082***	
	(0.021)	
internet_sm	0.050**	
	(0.020)	
discussions	0.066***	
	(0.024)	
podcasts	0.050**	
	(0.022)	
twitter		0.286***
		(0.100)
facebook		−0.020
		(0.102)
instagram		−0.142
		(0.105)
snapchat		0.100
		(0.114)
reddit		−0.079
		(0.103)
tiktok		0.110
		(0.123)
youtube		0.227*
		(0.128)
Constant	−0.792***	−0.480*

E Phrase Slants (Gammas)

I here plot the γ parameter estimates, which represent phrases' slant or discrimination in the speech ideology space. Phrases with large negative-signed gamma values are highly discriminative in the liberal direction (i.e. their usage is highly diagnostic of speech liberalism) and those with large positive-signed gammas are highly discriminative in the conservative direction. Phrases with near-zero slant are relatively uninformative of speech ideology in this model.

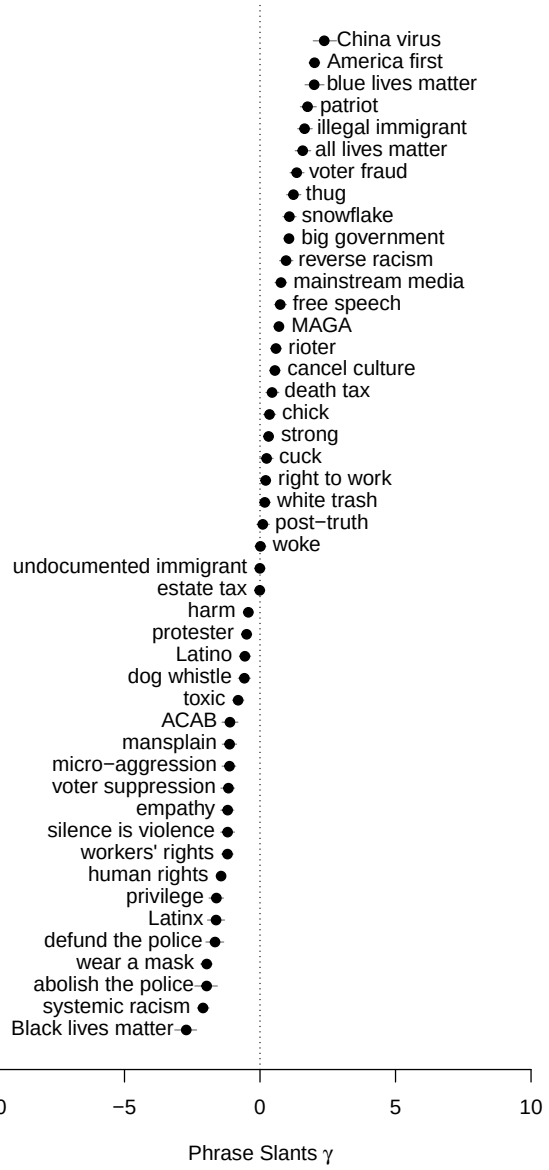

Figure 7: Gamma slant parameter estimates for all phrases.

F Bivariate Response Scatterplots

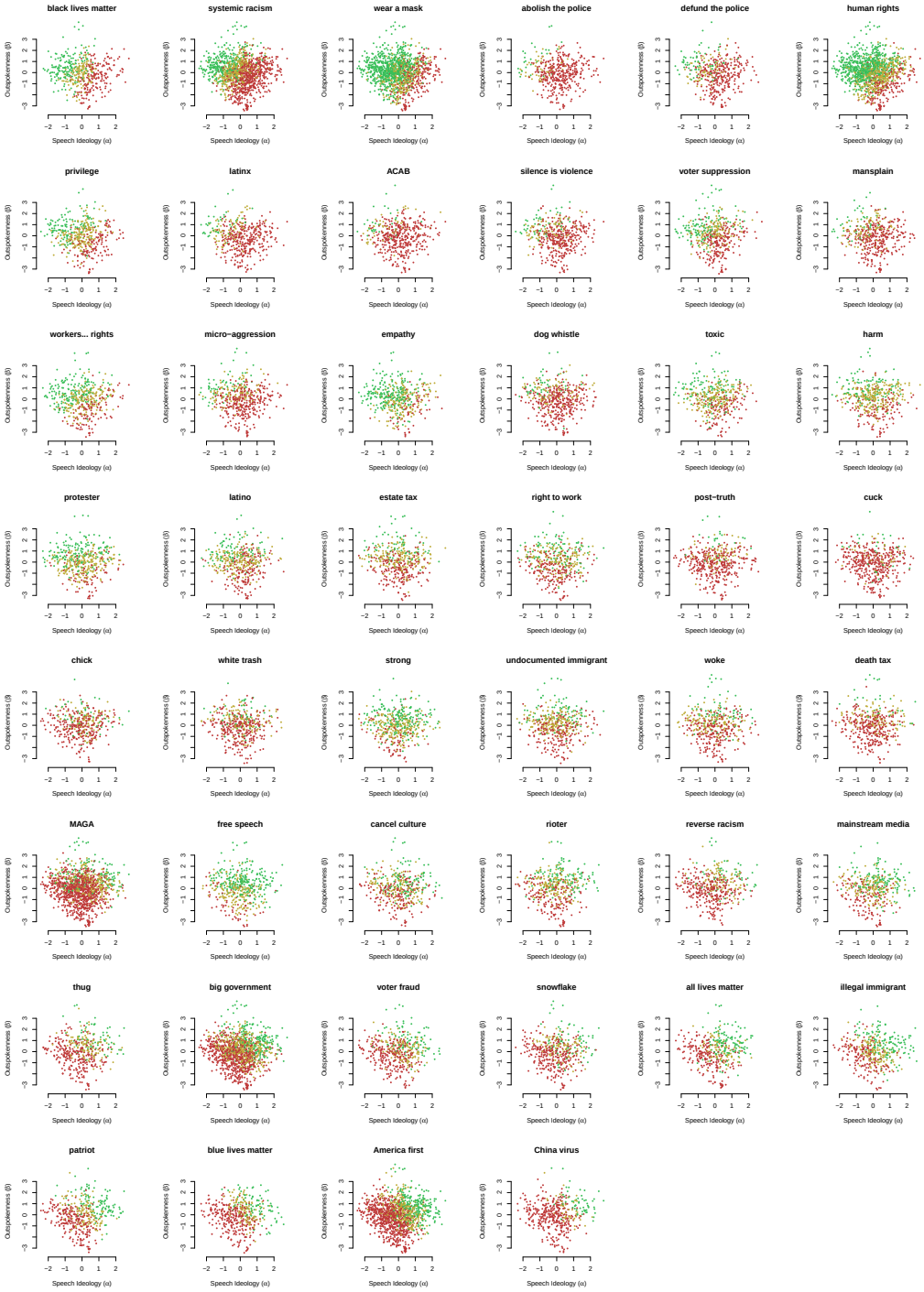

Figure 8: Bivariate response scatterplots (as in Figure 2) for all 46 phrases, arranged in order of γ slant.

G Stick Representations of Phrase Ideologies

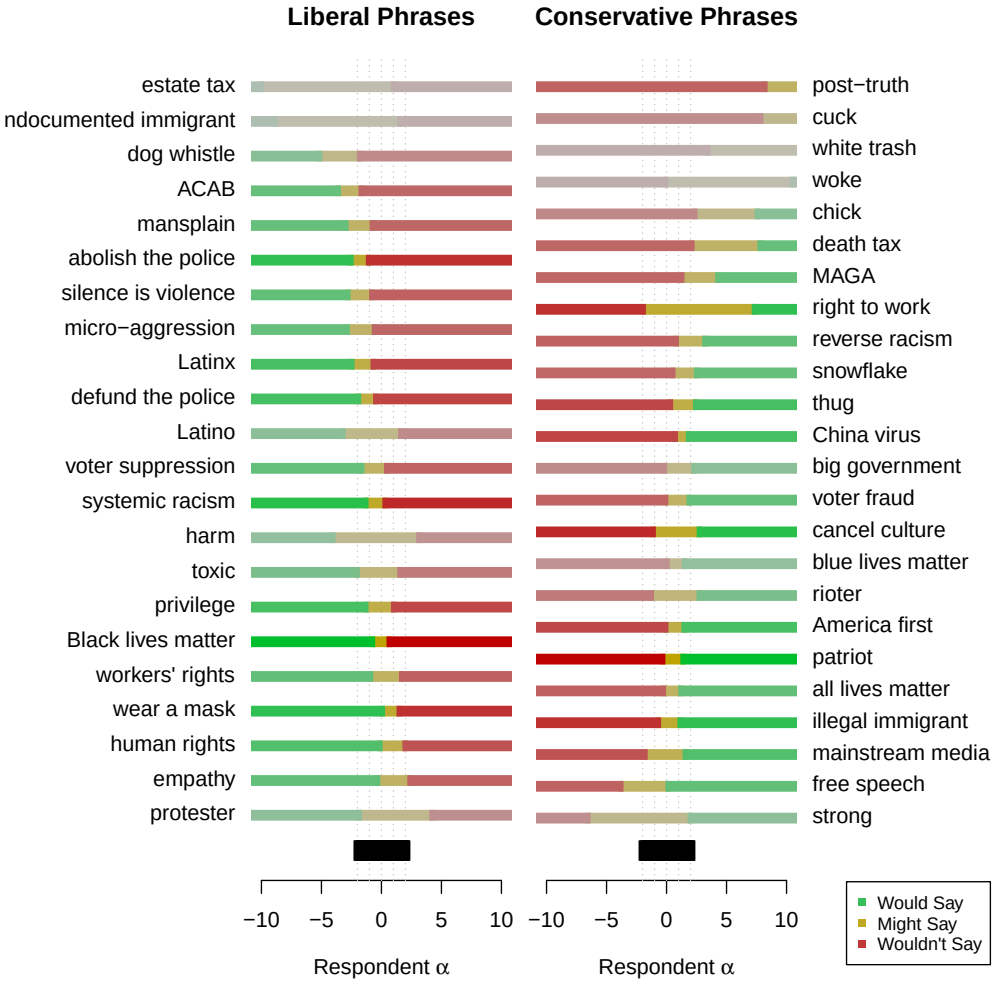

Figure 9: Predicted phrase usage response category as a function of respondent speech ideology α (as in Figure 3) for all 46 phrases. Note that color saturation is used to represent phrase absolute discrimination abs(γ): discriminative phrases are more vivid, while undiscriminative phrases are grayed out, reflecting the relative uninformativeness of respondent α in predicting response. This approach to scaling the phrases is less meaningful for undiscriminative phrases.

H Monotonicity Checks

The item response theory specification described in Appendix A and implemented in STAN code in Appendix B assumes that phrase usage is monotonic in respondent speech ideology. That is, the model is specified such that respondents are predicted to report greater (lesser) propensity to use right-slanted phrases, the more conservative (liberal) is their speech ideology, and vice-versa for left-slanted phrases.

Phrases in this plot are arranged (reading from left-to-right and top-to-bottom) in order of most negative slant/discrimination parameter γ to most positive γ, such that phrases with weak (near-zero) γ appear in the middle rows of the plot. Note the poor discrimination of these weakly-slanted phrases is reflected in the poor separation between the plotted densities.

Most phrases satisfy monotonicity, in that the yellow "might say this" density peaks between the green and red densities. However there appear to be several examples of phrases whose usage is non-monotonic with respect to speech ideology α. Specifically, "undocumented immigrant," "woke," "post-truth," "right to work," and "cuck," show signs of non-monotonicity.

Although these are a small minority of the phrases studied, it bears mentioning that the α trait has been estimated so as to fit a monotonic model to the data, and so α represents a dimension that predicts phrase usage in a monotonic fashion. A further robustness check might therefore estimate a spatial proximity model of phrase usage, in which usage declines as a respondent moves further away from the phrase ideal point in both directions.

A more elaborate extension would allow phrase usage to peak at multiple points along the latent speech ideology spectrum, however this model is beyond the scope of the present paper.

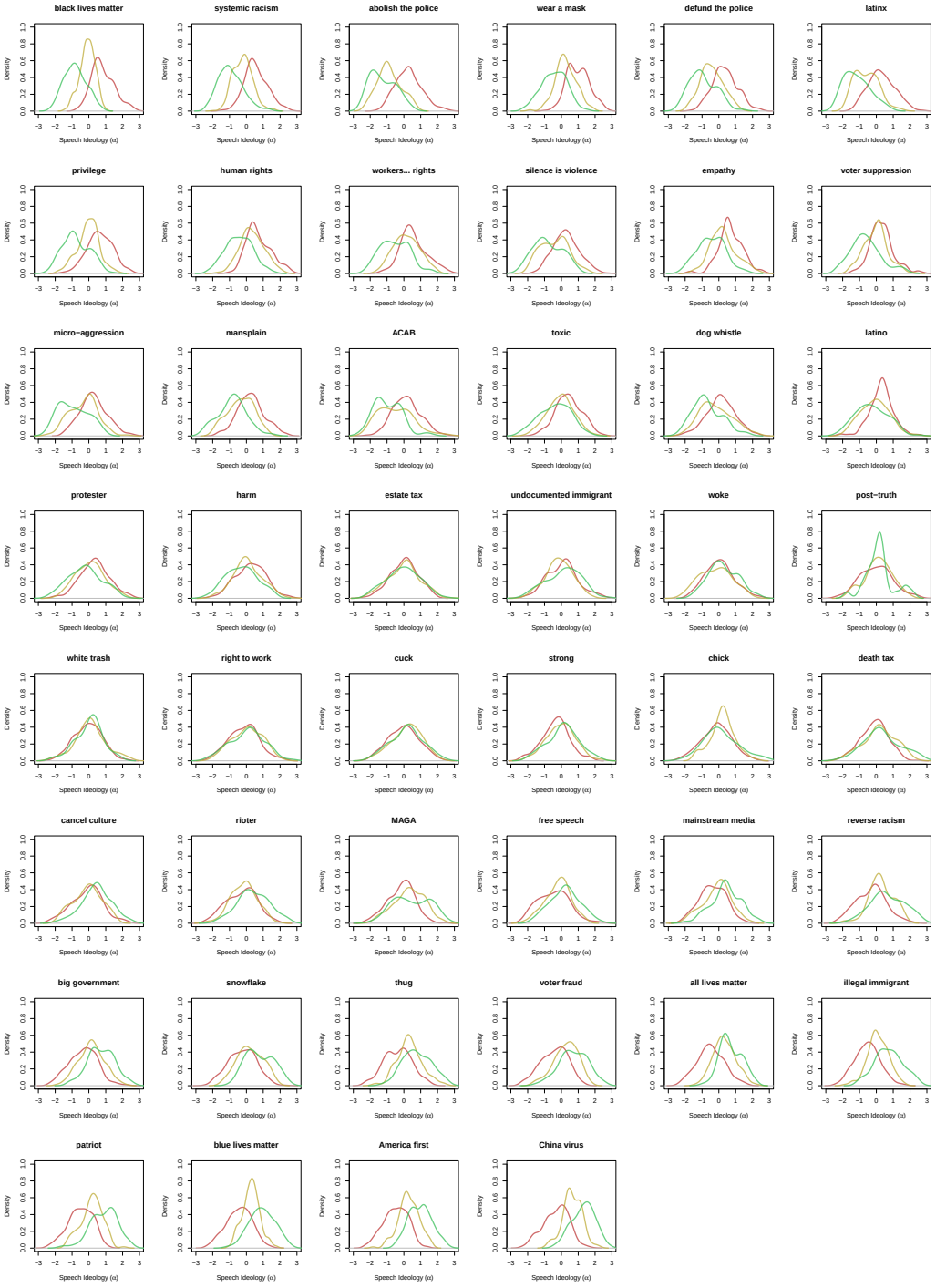

Figure 10: Density plots of observed responses, by respondent speech ideology α for all 46 phrases, arranged in order of γ slant.

I Results Using Wordfish Model

As stated in Section 4, the Wordsticks model that I introduce in this paper is simply an ordinal extension of the Wordfish model of text ideology that was introduced by Slapin and Proksch (2008). In this appendix, I present the results that follow from estimating a version of my model using the original Poisson specification from Slapin and Proksch. For ease of comparison to my Wordsticks model, I adapt the code for that model (as seen in Appendix B) to a Poisson specification while leaving fixed all other aspects of the model (note that the parameter `theta` here fills a similar role to the cutpoints `c` in the Wordsticks model):

```
data {
  int<lower=1> N;                     //number of data points
  int<lower=1> J;                     //number of subjects
  int<lower=1> K;                     //number of items
  int<lower=0,upper=2> say[N];        //outcome: "I would say this"/"I might say this"/"I would not say this"
  int<lower=1, upper=J> subj[N];      //subject id
  int<lower=1, upper=K> item[N];      //item id
}

parameters {
  vector[J] alpha_raw;         // respondent speech ideology (raw)
  vector[J] beta;              // respondent "outspokenness"
  vector[K] gamma;             // item slants (direction and strength of association with speech ideology)
  vector[K] theta;             // item "sayability" intercepts
  real mu_gamma;               // slant mean
  real<lower=0.1> sigma_gamma; // slant spread
  real<lower=0.1> sigma_beta;  // outspokenness spread
  real mu_theta;               // phrase intercept mean
  real<lower=0.1> sigma_theta; // phrase intercept spread
}

transformed parameters {
  vector[J] alpha;             // hard-standardize alphas to aid model identification
  alpha = (alpha_raw - mean(alpha_raw)) ./ sd(alpha_raw);
}

model {
  //priors
  gamma[6] ~ exponential(.1);                 // enforce America first on the right
  gamma ~ normal(mu_gamma,sigma_gamma);       // normal prior on slants
  theta ~ normal(mu_theta,sigma_theta);       // normal prior on thetas
  alpha_raw ~ normal(0,1);                    // standard normal prior on speech ideology
  beta ~ normal(0,sigma_beta);                // normal prior (mean zero) on statement slants

  for (i in 1:N){
    say[i] ~ poisson(exp(gamma[item[i]] * alpha[subj[i]] + beta[subj[i]] + theta[item[i]]));
  }
}
```

I now present figures and tables reporting the same analyses presented in Section 6, but using the estimates of α and β returned from this Wordfish model:

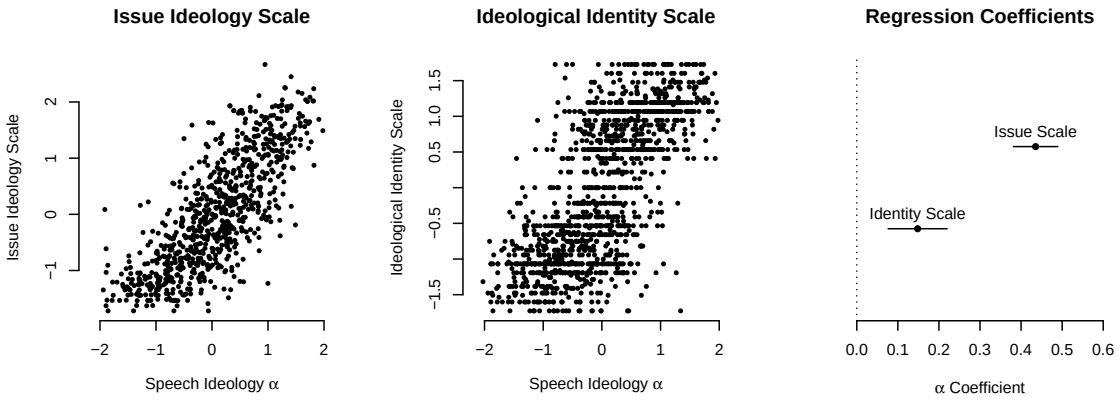

Figure 11: Comparison of speech ideology estimates (y axis) against issue ideology and signed identity strength, as in Figure 4. The right panel plots regression coefficients for both of these measures, in a model with respondent speech ideology α as the dependent variable (see Table 6, column 3).

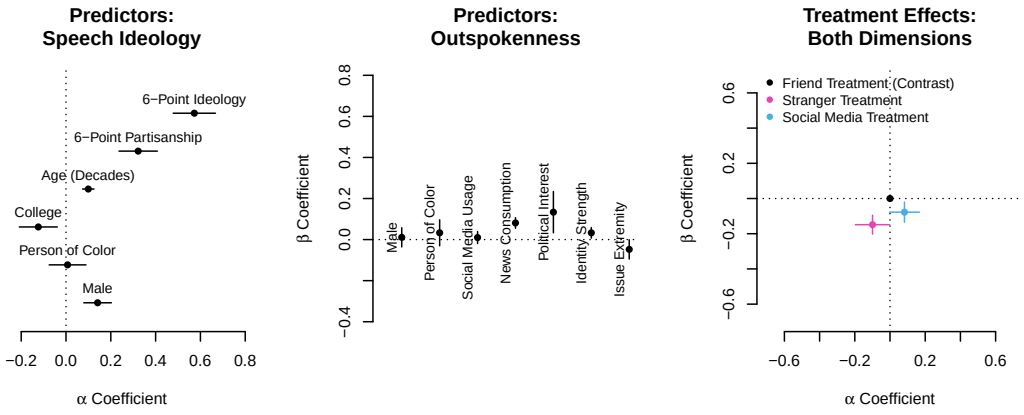

Figure 12: The left panel plots coefficients, with 95% confidence intervals, from a regression model with respondent speech ideology α as the dependent variable (see Table 6, column 5), using pooled data from both studies. The center panel plots coefficients, with 95% confidence intervals, from a model with respondent outspokenness β as the dependent variable (see Table 7, column 7). The right panel plots the effects of the "stranger" (pink) and "social media" (blue) treatments, relative to the common contrast of "close friend" (black, by construction at the origin) in both dimensions: the x axis denotes left-right shifts in speech ideology α (see Table 8, columns 2 and 4), and the y axis denotes up-down shifts in outspokenness β (see Table 9, columns 2 and 4).

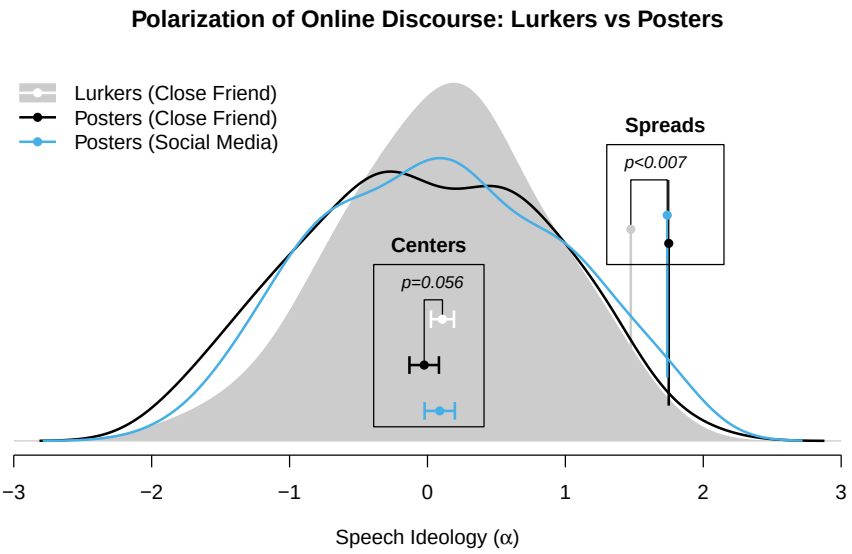

Figure 13: Density plots of respondent speech ideology α, in three exhaustive subsets of the Study 2 data: lurkers speaking with a close friend (gray), posters speaking with a close friend (black), and posters posting on their preferred online platform (blue).

As can be seen in the plots above and the tables below, using Wordfish to estimate α and β does not dramatically affect the key results of these analyses. The size and statistical significance of the descriptive predictors of α remains essentially the same. The spread of β is somewhat narrower, but the same variables – news consumption, political interest, and ideological identity strength – are identified as statistically significant predictors. The only notable difference is that, whereas the Wordsticks results indicated that the negative-signed treatment coefficient on speech ideology for the "stranger" context treatment was not significantly different from zero at conventional levels, in the Wordfish results this effect does barely reach conventional levels of statistical significance. Both the stranger and social media treatments are found to have a significant negative effect on outspokenness, as was found using the results from the Wordsticks model. Since the Wordfish model is incorrectly specified for the ordinal outcome, I assert that the lone difference in results – a significant negative effect on speech ideology of the stranger treatment – should be assumed to be spurious, although confirmatory evidence from a correctly-specified model might change this assessment.

Table 6: Alpha Descriptive Analyses

	Pilot 1 (Prolific)	Pilot 1 w/o SDI	Pilot 2 (MTurk)	Pilot 2 w/o Issues	Pooled w/o Identity	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
age_dec	0.101***	0.107***	0.084***	0.101***	0.100***	0.101***
	(0.022)	(0.022)	(0.014)	(0.013)	(0.013)	(0.013)
sdi_num6	0.510***			0.417***	0.573***	0.417***
	(0.115)			(0.072)	(0.048)	(0.072)
pid_num6	0.283***	0.393***	0.130**	0.282***	0.322***	0.282***
	(0.072)	(0.069)	(0.051)	(0.046)	(0.044)	(0.046)
pol_int	−0.050	−0.107	−0.092	−0.086	−0.070	−0.086
	(0.101)	(0.102)	(0.071)	(0.063)	(0.063)	(0.063)
issue_scale			0.435***			
			(0.028)			
signed_identity	0.093	0.336***	0.148***	0.133***		0.133***
	(0.074)	(0.050)	(0.037)	(0.046)		(0.046)
college	−0.204***	−0.206***	−0.007	−0.118***	−0.124***	−0.118***
	(0.069)	(0.071)	(0.049)	(0.043)	(0.043)	(0.043)
POC	0.021	0.034	−0.007	0.007	0.007	0.007
	(0.066)	(0.067)	(0.048)	(0.042)	(0.042)	(0.042)
male	0.145***	0.142***	0.078**	0.141***	0.141***	0.141***
	(0.051)	(0.052)	(0.035)	(0.031)	(0.031)	(0.031)
media_scale	−0.003	0.010	0.062***	0.038**	0.037**	0.038**
	(0.027)	(0.028)	(0.019)	(0.017)	(0.017)	(0.017)
sm_scale	−0.006	−0.007	0.014	−0.001	−0.003	−0.001
	(0.025)	(0.026)	(0.022)	(0.018)	(0.018)	(0.018)
Constant	−0.267**	−0.260**	−0.266***	−0.276***	−0.272***	−0.276***
	(0.118)	(0.120)	(0.085)	(0.073)	(0.074)	(0.073)
Observations	502	502	797	1,299	1,299	1,299
R^2	0.570	0.552	0.666	0.568	0.565	0.568
Adjusted R^2	0.561	0.544	0.662	0.564	0.562	0.564

Note: *p<0.1; **p<0.05; ***p<0.01

Table 7: Beta Descriptive Analyses

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
age_dec	0.004	0.009	0.023**	0.004	0.011	0.008	0.011
	(0.007)	(0.009)	(0.009)	(0.007)	(0.008)	(0.008)	(0.010)
sdi_num6	−0.005	−0.021	−0.016	−0.023	−0.001	−0.020	−0.027
	(0.029)	(0.037)	(0.039)	(0.028)	(0.029)	(0.028)	(0.037)
pid_num6	−0.055**	−0.062*	−0.079**	−0.043*	−0.055**	−0.043*	−0.059*
	(0.026)	(0.034)	(0.035)	(0.026)	(0.026)	(0.025)	(0.033)
pol_int	0.180***	0.261***		0.061	0.166***	0.055	0.134***
	(0.037)	(0.048)		(0.038)	(0.037)	(0.038)	(0.051)
identity_strength	0.049***	0.045***		0.040***	0.048***	0.040***	0.033**
	(0.010)	(0.013)		(0.010)	(0.010)	(0.010)	(0.013)
college	−0.006	−0.025	−0.009	−0.022	−0.005	−0.021	−0.028
	(0.026)	(0.034)	(0.035)	(0.025)	(0.026)	(0.025)	(0.033)
POC	0.065***	0.044	0.025	0.049**	0.061**	0.046*	0.033
	(0.025)	(0.033)	(0.034)	(0.024)	(0.025)	(0.024)	(0.032)
male	−0.001	0.013	0.022	−0.001	−0.0003	−0.001	0.011
	(0.019)	(0.024)	(0.025)	(0.019)	(0.019)	(0.019)	(0.024)
abs(issue_scale)		−0.075***	−0.009				−0.048**
		(0.024)	(0.023)				(0.024)
media_scale				0.090***		0.089***	0.081***
				(0.010)		(0.010)	(0.013)
sm_scale					0.025**	0.014	0.010
					(0.011)	(0.010)	(0.015)
Constant	−0.135***	−0.138**	−0.098*	−0.041	−0.154***	−0.054	−0.087
	(0.042)	(0.058)	(0.057)	(0.042)	(0.043)	(0.043)	(0.059)
Observations	1,299	797	797	1,299	1,299	1,299	797
R^2	0.077	0.100	0.038	0.133	0.081	0.134	0.143
Adjusted R^2	0.071	0.089	0.030	0.127	0.074	0.127	0.131

Note: *p<0.1; **p<0.05; ***p<0.01

Table 8: Alpha Treatment Effects

	(1)	(2)	(3)	(4)
age_dec		0.100***		0.085***
		(0.022)		(0.015)
sdi_num6		0.504***		0.010
		(0.115)		(0.085)
pid_num6		0.289***		0.126**
		(0.072)		(0.053)
pol_int		−0.047		−0.087
		(0.101)		(0.072)
issue_scale				0.434***
				(0.029)
signed_identity		0.092		0.144***
		(0.073)		(0.052)
college		−0.202***		−0.015
		(0.069)		(0.049)
POC		0.026		−0.012
		(0.066)		(0.048)
male		0.146***		0.077**
		(0.051)		(0.035)
media_scale		−0.005		0.066***
		(0.027)		(0.019)
sm_scale		−0.007		0.017
		(0.025)		(0.023)
stranger	−0.132*	−0.099**		
	(0.075)	(0.050)		
expressor			−0.133*	−0.067
			(0.070)	(0.043)
smct_X_expressor			0.113	0.082*
			(0.074)	(0.043)

Table 9: Beta Treatment Effects

	(1)	(2)	(3)	(4)	(5)
age_dec		0.007		0.010	0.004
		(0.012)		(0.010)	(0.012)
sdi_num6		−0.027		−0.030	−0.037
		(0.042)		(0.037)	(0.043)
pid_num6		0.005		−0.057*	−0.054
		(0.038)		(0.033)	(0.039)
pol_int		−0.028		0.114**	0.134**
		(0.060)		(0.050)	(0.062)
identity_strength		0.068***		0.029**	0.050***
		(0.016)		(0.013)	(0.016)
college		−0.010		−0.027	−0.025
		(0.038)		(0.033)	(0.039)
POC		0.059		0.041	0.081**
		(0.036)		(0.032)	(0.040)
male		−0.013		0.007	−0.015
		(0.029)		(0.024)	(0.029)
media_scale		0.089***		0.086***	0.071***
		(0.015)		(0.013)	(0.015)
sm_scale		0.021		0.012	0.013
		(0.014)		(0.015)	(0.018)
stranger	−0.158***	−0.149***			
	(0.029)	(0.027)			
expressor			0.090***	0.006	
			(0.029)	(0.029)	
smct_X_expressor			−0.088***	−0.078***	−0.085***
			(0.031)	(0.029)	(0.028)
sm_friend_likemindedness_diff					−0.038
					(0.043)
smct_X_expressor:sm_friend_likemindedness_diff					0.139**
					(0.060)
Constant	0.081***	0.087	−0.024	−0.093	−0.071
	(0.020)	(0.067)	(0.020)	(0.060)	(0.073)
	54				
Observations	503	502	798	797	498
R^2	0.055	0.201	0.014	0.148	0.192
Adjusted R^2	0.053	0.183	0.012	0.135	0.170

Note: *p<0.1; **p<0.05; ***p<0.01

J Results Simulating a Binary Response Scale

Readers may be interested to know why I included an intermediate "might say this" response option in the What Would You Say question. It is, after all, much simpler to estimate an ideal point model with a binary outcome than an ordinal outcome. To characterize the role of the "might" option in this analysis, I here simulate a binary response scale by recoding "might" responses as either "would" or "wouldn't" responses, and re-running the main analyses presented in Section 6 using estimates of α and β from a version of Wordsticks adapted to a binary outcome. As with the Poisson specification presented in Appendix I, I adapt the code for my Wordsticks model to the case of a binary outcome while leaving fixed all other aspects of the model (note that the parameter `theta` here fills a similar role to the cutpoints `c` in the Wordsticks model):

```
data {
  int<lower=1> N;                  //number of data points
  int<lower=1> J;                  //number of subjects
  int<lower=1> K;                  //number of items
  int<lower=0,upper=1> say[N];     //outcome: "I would say this"/"I would not say this"
  int<lower=1, upper=J> subj[N];   //subject id
  int<lower=1, upper=K> item[N];   //item id
}

parameters {
  vector[J] alpha_raw;         // respondent speech ideology (raw)
  vector[J] beta;              // respondent "outspokenness"
  vector[K] gamma;             // item slants (direction and strength of association with speech ideology)
  vector[K] c;                 // cutpoint location parameters
  real mu_gamma;               // slant mean
  real<lower=0.1> sigma_gamma; // slant spread
  real<lower=0.1> sigma_beta;  // outspokenness spread
}

transformed parameters { // hard-standardize alphas to aid model identification
  vector[J] alpha;
  alpha = (alpha_raw - mean(alpha_raw)) ./ sd(alpha_raw);
}

model {
  //priors
  gamma[6] ~ exponential(.1);            // enforce America first right-slanted
  gamma ~ normal(mu_gamma,sigma_gamma);  // normal prior on slants
  alpha_raw ~ normal(0,1);               // standard normal prior on speech ideology
  beta ~ normal(0,sigma_beta);           // normal prior (mean zero) on statement slants

  for (i in 1:N){
    say[i] ~ bernoulli_logit(gamma[item[i]] * alpha[subj[i]] + beta[subj[i]] - c[item[i]]);
  }
}
```

I now present figures and tables reporting the same analyses presented in Section 6, but using the estimates of α and β returned from the binary model, under two alternative simulations of a binary response scale: one in which I convert "I might say this" response to "I wouldn't say this" responses, and another in which I convert "might" to "would."

J.1 Might to Wouldn't

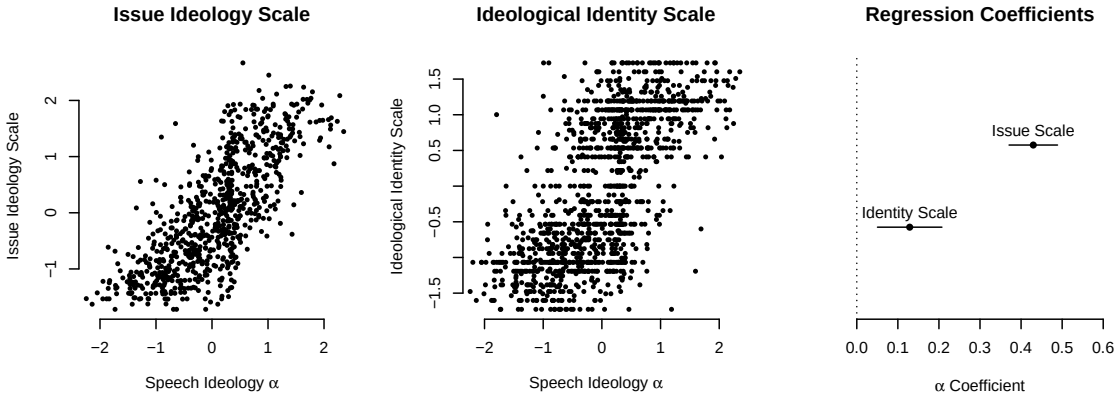

Figure 14: Comparison of speech ideology estimates (y axis) against issue ideology and signed identity strength, as in Figure 4. The right panel plots regression coefficients for both of these measures, in a model with respondent speech ideology α as the dependent variable (see Table **??**, column 3).

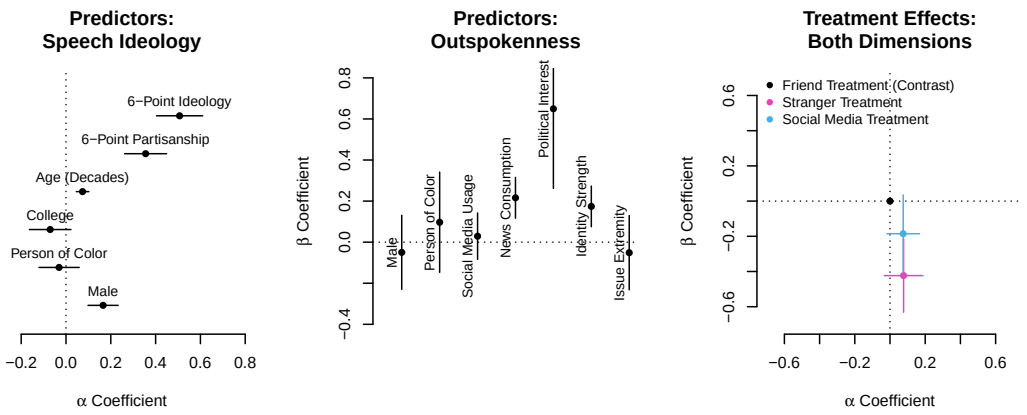

Figure 15: The left panel plots coefficients, with 95% confidence intervals, from a regression model with respondent speech ideology α as the dependent variable (see Table **??**, column 5), using pooled data from both studies. The center panel plots coefficients, with 95% confidence intervals, from a model with respondent outspokenness β as the dependent variable (see Table **??**, column 7). The right panel plots the effects of the "stranger" (pink) and "social media" (blue) treatments, relative to the common contrast of "close friend" (black, by construction at the origin) in both dimensions: the x axis denotes left-right shifts in speech ideology α (see Table **??**, columns 2 and 4), and the y axis denotes up-down shifts in outspokenness β (see Table **??**, columns 2 and 4).

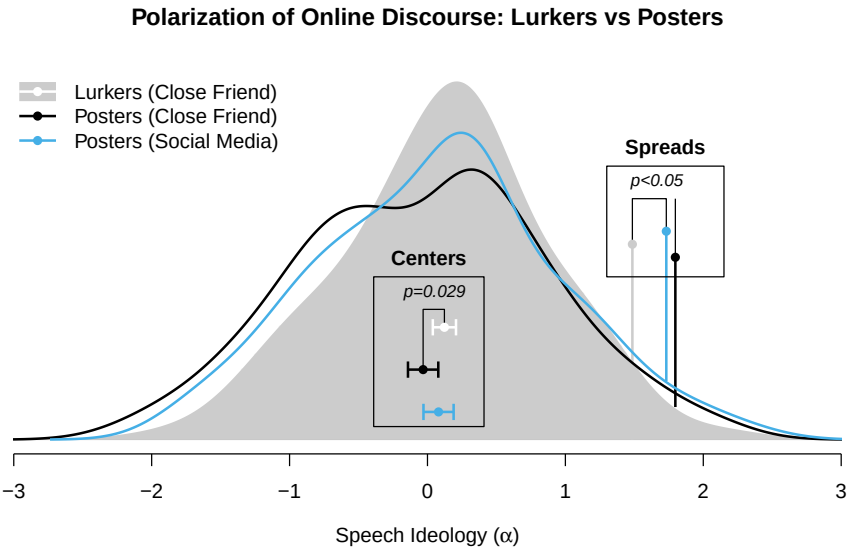

Figure 16: Density plots of respondent speech ideology α, in three exhaustive subsets of the Study 2 data: lurkers speaking with a close friend (gray), posters speaking with a close friend (black), and posters posting on their preferred online platform (blue).

Table 10: Alpha Descriptive Analyses

	Pilot 1 (Prolific)	Pilot 1 w/o SDI	Pilot 2 (MTurk)	Pilot 2 w/o Issues	Pooled w/o Identity	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
age_dec	0.079***	0.083***	0.055***	0.075***	0.074***	0.075***
	(0.025)	(0.025)	(0.016)	(0.014)	(0.014)	(0.014)
sdi_num6	0.342***			0.327***	0.507***	0.327***
	(0.130)			(0.078)	(0.053)	(0.078)
pid_num6	0.308***	0.382***	0.155***	0.309***	0.355***	0.309***
	(0.081)	(0.077)	(0.056)	(0.050)	(0.048)	(0.050)
pol_int	−0.037	−0.076	−0.110	−0.089	−0.071	−0.089
	(0.114)	(0.113)	(0.078)	(0.069)	(0.069)	(0.069)
issue_scale			0.430***			
			(0.030)			
signed_identity	0.148*	0.311***	0.129***	0.154***		0.154***
	(0.083)	(0.055)	(0.040)	(0.050)		(0.050)
college	−0.129*	−0.131*	0.033	−0.064	−0.071	−0.064
	(0.078)	(0.078)	(0.054)	(0.047)	(0.047)	(0.047)
POC	0.042	0.051	−0.080	−0.031	−0.031	−0.031
	(0.074)	(0.075)	(0.052)	(0.046)	(0.046)	(0.046)
male	0.163***	0.162***	0.108***	0.166***	0.166***	0.166***
	(0.058)	(0.058)	(0.038)	(0.034)	(0.034)	(0.034)
media_scale	−0.058*	−0.049	0.016	−0.011	−0.012	−0.011
	(0.031)	(0.031)	(0.021)	(0.019)	(0.019)	(0.019)
sm_scale	−0.032	−0.033	0.021	−0.009	−0.012	−0.009
	(0.028)	(0.028)	(0.024)	(0.019)	(0.019)	(0.019)
Constant	−0.278**	−0.273**	−0.162*	−0.222***	−0.217***	−0.222***
	(0.133)	(0.133)	(0.092)	(0.080)	(0.080)	(0.080)
Observations	502	502	797	1,299	1,299	1,299
R^2	0.480	0.473	0.611	0.500	0.496	0.500
Adjusted R^2	0.470	0.464	0.606	0.496	0.493	0.496

Note: *p<0.1; **p<0.05; ***p<0.01

Table 11: Beta Descriptive Analyses

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
age_dec	0.004	0.013	0.072**	0.002	0.025	0.017	0.019
	(0.028)	(0.036)	(0.036)	(0.027)	(0.030)	(0.029)	(0.038)
sdi_num6	0.113	0.044	0.070	0.068	0.125	0.077	0.031
	(0.108)	(0.143)	(0.148)	(0.107)	(0.109)	(0.108)	(0.141)
pid_num6	−0.144	−0.164	−0.236*	−0.113	−0.143	−0.113	−0.156
	(0.099)	(0.130)	(0.134)	(0.098)	(0.099)	(0.098)	(0.129)
pol_int	0.819***	0.990***		0.520***	0.774***	0.498***	0.649***
	(0.139)	(0.183)		(0.147)	(0.141)	(0.147)	(0.198)
identity_strength	0.238***	0.206***		0.218***	0.236***	0.216***	0.174***
	(0.038)	(0.051)		(0.038)	(0.038)	(0.038)	(0.051)
college	−0.056	−0.057	0.003	−0.096	−0.051	−0.092	−0.065
	(0.098)	(0.129)	(0.133)	(0.097)	(0.098)	(0.097)	(0.127)
POC	0.194**	0.127	0.053	0.154	0.181*	0.146	0.097
	(0.094)	(0.126)	(0.130)	(0.093)	(0.095)	(0.094)	(0.125)
male	−0.079	−0.043	−0.020	−0.079	−0.076	−0.077	−0.050
	(0.072)	(0.093)	(0.094)	(0.071)	(0.072)	(0.071)	(0.092)
abs(issue_scale)		−0.124	0.148*				−0.051
		(0.092)	(0.088)				(0.092)
media_scale				0.225***		0.220***	0.216***
				(0.038)		(0.038)	(0.051)
sm_scale					0.078**	0.052	0.029
					(0.040)	(0.039)	(0.058)
Constant	−0.445***	−0.534**	−0.407*	−0.212	−0.506***	−0.259	−0.398*
	(0.160)	(0.221)	(0.216)	(0.163)	(0.163)	(0.167)	(0.227)
Observations	1,299	797	797	1,299	1,299	1,299	797
R^2	0.092	0.087	0.016	0.116	0.095	0.117	0.109
Adjusted R^2	0.086	0.077	0.008	0.110	0.088	0.111	0.096

Note: *p<0.1; **p<0.05; ***p<0.01

Table 12: Alpha Treatment Effects

	(1)	(2)	(3)	(4)
age_dec		0.080***		0.056***
		(0.025)		(0.016)
sdi_num6		0.346***		−0.024
		(0.130)		(0.093)
pid_num6		0.303***		0.157***
		(0.081)		(0.058)
pol_int		−0.039		−0.106
		(0.114)		(0.078)
issue_scale				0.431***
				(0.031)
signed_identity		0.149*		0.139**
		(0.083)		(0.057)
college		−0.131*		0.027
		(0.078)		(0.054)
POC		0.038		−0.085
		(0.074)		(0.052)
male		0.162***		0.107***
		(0.058)		(0.038)
media_scale		−0.056*		0.020
		(0.031)		(0.021)
sm_scale		−0.032		0.025
		(0.028)		(0.025)
stranger	0.056	0.077		
	(0.077)	(0.056)		
expressor			−0.154**	−0.067
			(0.071)	(0.047)
smct_X_expressor			0.112	0.075
			(0.075)	(0.047)

Table 13: Beta Treatment Effects

	(1)	(2)	(3)	(4)	(5)
age_dec		0.034		0.017	−0.006
		(0.047)		(0.038)	(0.048)
sdi_num6		0.106		0.021	−0.049
		(0.164)		(0.141)	(0.172)
pid_num6		0.014		−0.147	−0.080
		(0.149)		(0.129)	(0.155)
pol_int		0.301		0.628***	0.792***
		(0.231)		(0.192)	(0.250)
identity_strength		0.309***		0.172***	0.263***
		(0.061)		(0.049)	(0.063)
college		−0.134		−0.057	−0.036
		(0.148)		(0.128)	(0.158)
POC		0.206		0.111	0.308*
		(0.141)		(0.124)	(0.160)
male		−0.119		−0.052	−0.129
		(0.111)		(0.091)	(0.114)
media_scale		0.214***		0.220***	0.162***
		(0.059)		(0.050)	(0.061)
sm_scale		0.086		0.030	0.017
		(0.054)		(0.059)	(0.073)
stranger	−0.462***	−0.423***			
	(0.114)	(0.106)			
expressor			0.353***	0.061	
			(0.111)	(0.112)	
smct_X_expressor			−0.205*	−0.185*	−0.220*
			(0.117)	(0.113)	(0.112)
sm_friend_likemindedness_diff					−0.214
					(0.172)
smct_X_expressor:sm_friend_likemindedness_diff					0.600**
					(0.239)
Constant	0.253***	0.134	−0.120	−0.412*	−0.366
	(0.079)	(0.260)	(0.075)	(0.230)	(0.292)
Observations	503	502	798	797	498
R^2	0.032	0.181	0.013	0.112	0.144

J.2 Might to Would

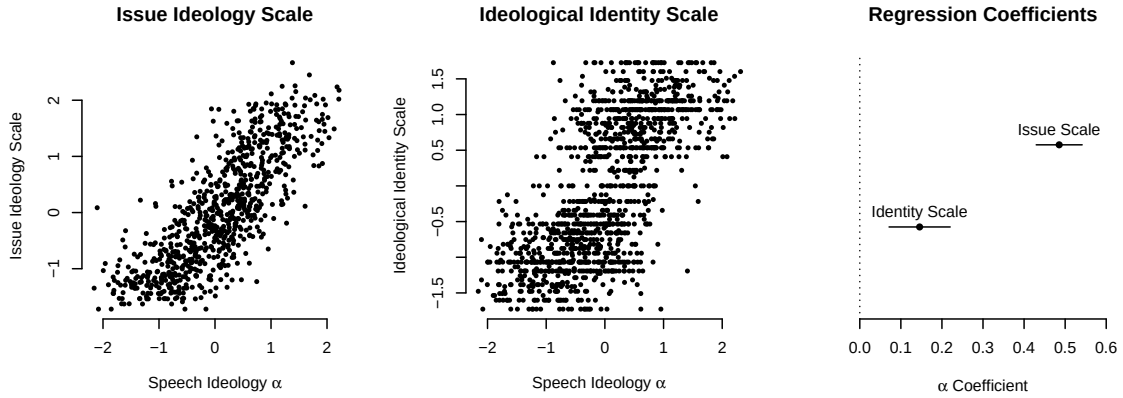

Figure 17: Comparison of speech ideology estimates (y axis) against issue ideology and signed identity strength, as in Figure 4. The right panel plots regression coefficients for both of these measures, in a model with respondent speech ideology α as the dependent variable (see Table **??**, column 3).

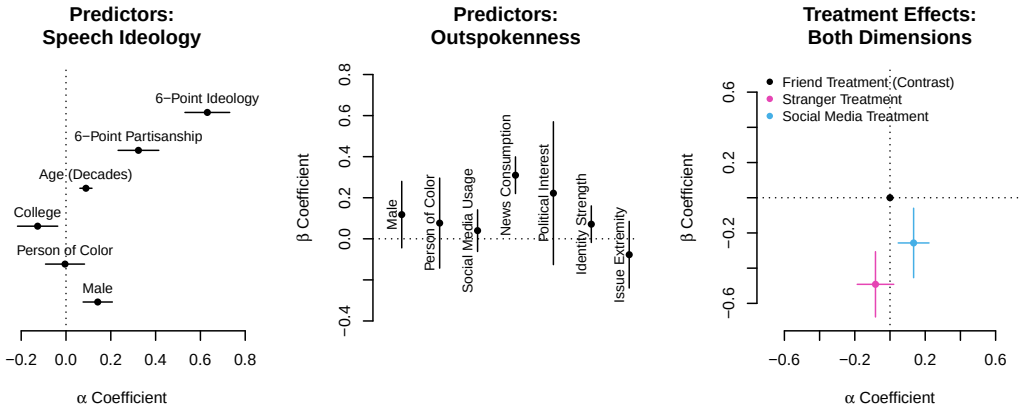

Figure 18: The left panel plots coefficients, with 95% confidence intervals, from a regression model with respondent speech ideology α as the dependent variable (see Table **??**, column 5), using pooled data from both studies. The center panel plots coefficients, with 95% confidence intervals, from a model with respondent outspokenness β as the dependent variable (see Table **??**, column 7). The right panel plots the effects of the "stranger" (pink) and "social media" (blue) treatments, relative to the common contrast of "close friend" (black, by construction at the origin) in both dimensions: the x axis denotes left-right shifts in speech ideology α (see Table **??**, columns 2 and 4), and the y axis denotes up-down shifts in outspokenness β (see Table **??**, columns 2 and 4).

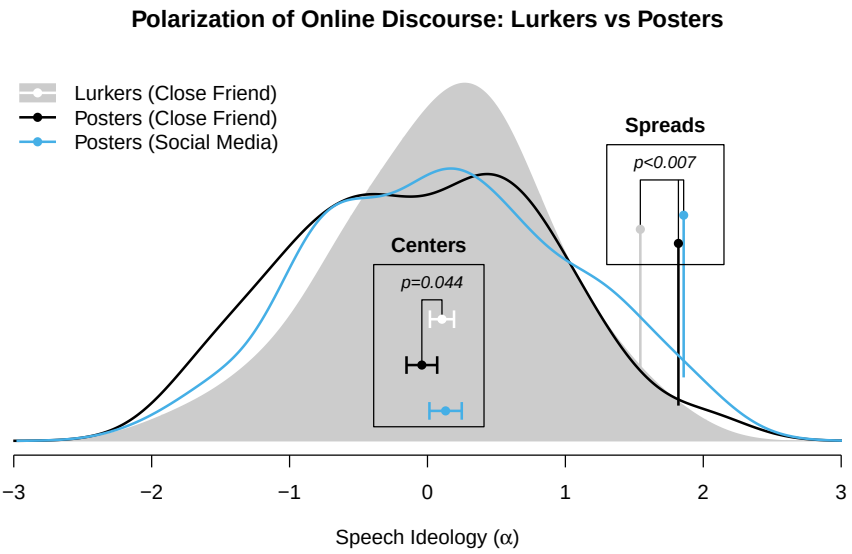

Figure 19: Density plots of respondent speech ideology α, in three exhaustive subsets of the Study 2 data: lurkers speaking with a close friend (gray), posters speaking with a close friend (black), and posters posting on their preferred online platform (blue).

J.3 Simulation Outcome Summary

To summarize the results of these simulations of a binary response scale, it would appear that the inclusion of an intermediate "might say this" response option is not absolutely necessary for measuring speech ideology, but is useful for measuring outspokenness. This conclusion is based on the findings that when "might" responses are converted to "wouldn't" responses, the effect of the social media context treatment on outspokenness loses significance, and that when "might" responses are converted to "would" responses, although this treatment effect remains, the descriptive coefficients on political interest and identity strength in predicting outspokenness lose significance. Although it is unknown which of these simulations most closely approximates the data that would have been collected under the counterfactual conditions of a binary "would" or "wouldn't" response scale, it seems fair to say that there are benefits to including a "might" option, since without it we might lose valuable information about either the descriptive covariates that characterize the β trait, or about the causal effects of the social context treatments, or possibly both.

Table 14: Alpha Descriptive Analyses

	Pilot 1 (Prolific)	Pilot 1 w/o SDI	Pilot 2 (MTurk)	Pilot 2 w/o Issues	Pooled w/o Identity	Pooled
	(1)	(2)	(3)	(4)	(5)	(6)
age_dec	0.089***	0.096***	0.071***	0.090***	0.089***	0.090***
	(0.023)	(0.024)	(0.015)	(0.014)	(0.014)	(0.014)
sdi_num6	0.663***			0.492***	0.631***	0.492***
	(0.123)			(0.076)	(0.051)	(0.076)
pid_num6	0.257***	0.400***	0.135**	0.288***	0.323***	0.288***
	(0.077)	(0.074)	(0.053)	(0.048)	(0.046)	(0.048)
pol_int	−0.085	−0.160	−0.143*	−0.133**	−0.118*	−0.133**
	(0.108)	(0.110)	(0.074)	(0.067)	(0.066)	(0.067)
issue_scale			0.485***			
			(0.029)			
signed_identity	0.044	0.360***	0.146***	0.119**		0.119**
	(0.078)	(0.053)	(0.038)	(0.048)		(0.048)
college	−0.224***	−0.227***	0.008	−0.122***	−0.127***	−0.122***
	(0.074)	(0.076)	(0.051)	(0.046)	(0.046)	(0.046)
POC	−0.003	0.014	−0.013	−0.004	−0.004	−0.004
	(0.070)	(0.072)	(0.049)	(0.044)	(0.044)	(0.044)
male	0.156***	0.153***	0.068*	0.142***	0.142***	0.142***
	(0.055)	(0.056)	(0.036)	(0.033)	(0.033)	(0.033)
media_scale	−0.031	−0.014	0.038*	0.013	0.012	0.013
	(0.029)	(0.030)	(0.020)	(0.018)	(0.018)	(0.018)
sm_scale	−0.004	−0.006	0.009	−0.001	−0.003	−0.001
	(0.027)	(0.028)	(0.023)	(0.019)	(0.019)	(0.019)
Constant	−0.187	−0.177	−0.173**	−0.193**	−0.190**	−0.193**
	(0.126)	(0.129)	(0.088)	(0.078)	(0.078)	(0.078)
Observations	502	502	797	1,299	1,299	1,299
R^2	0.561	0.535	0.677	0.562	0.560	0.562
Adjusted R^2	0.552	0.526	0.673	0.558	0.557	0.558

Note: *p<0.1; **p<0.05; ***p<0.01

Table 15: Beta Descriptive Analyses

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
age_dec	0.024	0.042	0.082**	0.022	0.050*	0.038	0.051
	(0.025)	(0.033)	(0.032)	(0.024)	(0.027)	(0.026)	(0.034)
sdi_num6	0.042	−0.012	0.004	−0.021	0.057	−0.011	−0.031
	(0.099)	(0.130)	(0.133)	(0.096)	(0.099)	(0.097)	(0.127)
pid_num6	−0.088	−0.095	−0.141	−0.043	−0.087	−0.044	−0.083
	(0.091)	(0.119)	(0.121)	(0.088)	(0.090)	(0.088)	(0.115)
pol_int	0.478***	0.710***		0.051	0.421***	0.025	0.222
	(0.128)	(0.168)		(0.132)	(0.129)	(0.132)	(0.177)
identity_strength	0.143***	0.116**		0.114***	0.140***	0.112***	0.071
	(0.035)	(0.046)		(0.034)	(0.035)	(0.034)	(0.045)
college	−0.022	−0.100	−0.058	−0.079	−0.016	−0.074	−0.113
	(0.090)	(0.118)	(0.119)	(0.087)	(0.089)	(0.087)	(0.114)
POC	0.193**	0.119	0.068	0.135	0.177**	0.126	0.077
	(0.087)	(0.115)	(0.117)	(0.084)	(0.087)	(0.084)	(0.112)
male	0.070	0.127	0.151*	0.070	0.074	0.072	0.118
	(0.066)	(0.085)	(0.085)	(0.064)	(0.066)	(0.064)	(0.083)
abs(issue_scale)		−0.181**	−0.006				−0.077
		(0.084)	(0.079)				(0.083)
media_scale				0.321***		0.315***	0.310***
				(0.034)		(0.034)	(0.046)
sm_scale					0.098***	0.060*	0.040
					(0.036)	(0.035)	(0.052)
Constant	−0.449***	−0.487**	−0.372*	−0.116	−0.525***	−0.170	−0.290
	(0.147)	(0.202)	(0.194)	(0.146)	(0.149)	(0.150)	(0.204)
Observations	1,299	797	797	1,299	1,299	1,299	797
R^2	0.045	0.054	0.016	0.107	0.050	0.109	0.109
Adjusted R^2	0.039	0.043	0.007	0.101	0.044	0.102	0.097

Note: *p<0.1; **p<0.05; ***p<0.01

Table 16: Alpha Treatment Effects

	(1)	(2)	(3)	(4)
age_dec		0.088***		0.071***
		(0.023)		(0.015)
sdi_num6		0.658***		−0.007
		(0.123)		(0.087)
pid_num6		0.262***		0.131**
		(0.077)		(0.055)
pol_int		−0.084		−0.153**
		(0.108)		(0.074)
issue_scale				0.486***
				(0.030)
signed_identity		0.043		0.150***
		(0.078)		(0.053)
college		−0.222***		0.001
		(0.074)		(0.051)
POC		0.002		−0.019
		(0.070)		(0.049)
male		0.157***		0.067*
		(0.055)		(0.036)
media_scale		−0.033		0.039**
		(0.029)		(0.020)
sm_scale		−0.005		0.009
		(0.027)		(0.023)
stranger	−0.112	−0.082		
	(0.079)	(0.053)		
expressor			−0.146**	−0.045
			(0.074)	(0.044)
		66		
smct_X_expressor			0.172**	0.134***
			(0.078)	(0.045)
Constant	0.034	0.148	0.195**	0.176**

Table 17: Beta Treatment Effects

	(1)	(2)	(3)	(4)	(5)
age_dec		0.024		0.049	0.031
		(0.042)		(0.034)	(0.041)
sdi_num6		−0.034		−0.042	−0.034
		(0.146)		(0.127)	(0.149)
pid_num6		0.079		−0.073	−0.097
		(0.132)		(0.115)	(0.134)
pol_int		−0.231		0.207	0.247
		(0.206)		(0.172)	(0.216)
identity_strength		0.211***		0.075*	0.137**
		(0.054)		(0.044)	(0.054)
college		0.003		−0.108	−0.104
		(0.132)		(0.114)	(0.137)
POC		0.190		0.094	0.194
		(0.126)		(0.111)	(0.138)
male		0.009		0.115	0.026
		(0.099)		(0.082)	(0.099)
media_scale		0.297***		0.322***	0.278***
		(0.052)		(0.045)	(0.052)
sm_scale		0.083*		0.052	0.066
		(0.048)		(0.052)	(0.064)
stranger	−0.522***	−0.492***			
	(0.100)	(0.095)			
expressor			0.240**	−0.019	
			(0.100)	(0.100)	
smct_X_expressor			−0.290***	−0.257**	−0.274***
			(0.106)	(0.101)	(0.097)
sm_friend_likemindedness_diff					−0.050
					(0.148)
smct_X_expressor:sm_friend_likemindedness_diff					0.387*
					(0.206)
Constant	0.231***	0.276	−0.046	−0.255	−0.204
	(0.070)	(0.232)	(0.068)	(0.205)	(0.252)
Observations	503	502	798	797	498
R^2	0.052	0.175	0.011	0.119	0.148
Adjusted R^2	0.050	0.156	0.009	0.105	0.125

K Issue Ideology Questions

To measure issue preferences on a variety of salient political topics, the following issue prompts were used in Study 2, with a 5-point agree-disagree response scale:

- Gun control laws in the United States should be stricter.
- Free trade agreements like the North American Free Trade Agreement (NAFTA) have helped the U.S. economy.
- A zero-tolerance policy for sexual harassment is essential to bringing about change in our society.
- Government regulation of business is necessary to protect the public interest.
- The U.S. should primarily take care of its own interests and let other countries get along the best they can on their own.
- Poor people today have it easy because they can get government benefits without doing anything in return.
- Business corporations make too much profit.
- Global warming will pose a serious threat to me or my way of life in my lifetime.
- Some police funding should be reallocated to other social services.
- There's too much pressure on Americans to get a COVID-19 vaccine.

L Ideological Identity Strength Questions

The following questions (adapted from Huddy, Mason and Aarøe, 2015) were used to measure respondents' strength of identification with liberalism or conservatism: respondents had previously indicated their perception of their own ideological position on a 5-point scale from "very conservative" to "very liberal," and those who chose the intermediate response, "moderate" were asked whether they leaned liberal or conservative. The appropriate label, "liberal" or "conservative," was then piped into the "identity_name" field in the question prompts below, so that the questions asked how strongly the respondents identified with their chosen ideological label.

How important is being a ${e://Field/identity_name} to you?

- Extremely important
- Very important
- Not very important
- Not important at all

How well does the term ${e://Field/identity_name} describe you?

- Extremely well
- Very well
- Not very well
- Not at all

When talking about ${e://Field/identity_name}s, how often do you use "we" instead of "they"?

- All of the time
- Most of the time
- Some of the time
- Rarely
- Never