Political Analysis (2025), **33**, 122–139
doi:10.1017/pan.2024.15

PA

ARTICLE

What Would You Say? Estimating Causal Effects of Social Context on Political Expression

William Small Schulz

Department of Politics, Princeton University, Princeton, NJ, USA. Email: wschulz@princeton.edu.

(Received 28 June 2023; revised 14 March 2024; accepted 16 March 2024; published online 27 December 2024)

Abstract

I develop a survey method for estimating social influence over individual political expression, by combining the content-richness of document scaling with the flexibility of survey research. I introduce the "What Would You Say?" question, which measures self-reported usage of political catchphrases in a hypothetical social context, which I manipulate in a between-subjects experiment. Using Wordsticks, an ordinal item response theory model inspired by Wordfish, I estimate each respondent's lexical ideology and outspokenness, scaling their political lexicon in a two-dimensional space. I then identify self-censorship and preference falsification as causal effects of social context on respondents' outspokenness and lexical ideology, respectively. This improves upon existing survey measures of political expression: it avoids conflating expressive behavior with populist attitudes, it defines preference falsification in terms of code-switching, and it moves beyond trait measures of self-censorship, to characterize relative shifts in the content of expression between different contexts. I validate the method and present experiments demonstrating its application to contemporary concerns about self-censorship and polarization, and I conclude by discussing its interpretation and future uses.

Keywords: speech; self-censorship; preference falsification; ideology; scaling; surveys; experiments

Edited by: Jeff Gill

1. Introduction

Even when speech is legally free, it is subject to intense social regulation. This is why, more than two centuries after the First Amendment was ratified in the United States, freedom of speech remains a *cause célèbre* today—particularly to those wary of "a new set of moral attitudes and political commitments that tend to weaken our norms of open debate and toleration of differences in favor of ideological conformity" (Ackerman *et al.* 2020).

Scholars have long worried that social constraints on speech may undermine a healthy polity by encouraging self-censorship (abstention from expressing opinions one believes are unpopular; see Noelle-Neumann 1974) and preference falsification (feigning opinions one believes are popular; see Kuran 1995). These theories originally sought to explain the appearance of social consensus under repressive regimes, but today they arise more often in the context of polarization and democratic dysfunction. For example, Sunstein (2019) argues that when citizens are surrounded by ideological compatriots, they tend to censor dissonant opinions and feign agreement with their "tribes," leading to cascading entrenchment and the polarization of political discourse.

A broader set of worries concern the idea that modern society is characterized by a "cancel culture" (Norris 2023) in which heterodox viewpoints are socially ostracized, both in academic communities (Revers and Traunmüller 2020) and in the general public (Menzner and Traunmüller 2023), and

(perhaps especially) on social media platforms (Weeks, Halversen, and Neubaum 2023). Many consider this anathema to the free exchange of ideas and related democratic ideals (e.g. Habermas 1989). For example, Chong, Citrin, and Levy (2022) document Americans' shrinking tolerance for speech deemed critical of identity groups, arguing that while this may reduce "the use of slurs and epithets . . . it has also significantly reduced citizens' freedom to deliberate about the political issues of the day" (19).

Although these concerns are widespread, they are difficult to investigate satisfactorily with existing survey methods. One problem is that "freedom of speech" is a politicized concept, which makes it hard to interpret responses to direct questions like, "Do you or don't you feel as free to speak your mind as you used to?" (Stouffer 1955). This question has been posed to representative samples over a long time period, so it has virtues for longitudinal analyses like those conducted by Gibson and Sutherland (2023), who find these feelings of unfreedom have increased over time—particularly among conservatives. However, what this question measures is a *subjective perception of a loss* of free expression, and since this perception has become "a constitutive element of populist ideology" (Menzner and Traunmüller 2023, 176), the resulting data may reflect respondents' populist attitudes more than their behavioral self-censorship. Such a charged topic requires a more circumspect approach to data collection.

Self-censorship is inherently difficult to observe because it is a non-event, like Holmes' "dog that did not bark" (Norris 2023). Although there exist well-validated scales of "willingness-to-self-censor" (e.g., Hayes, Uldall, and Glynn 2010), these are intended to measure self-censorship as a stable trait of an individual, rather than a behavior induced by certain social situations, and do not tell us which specific views are censored. List experiments (e.g., Blair and Imai 2012; Kuklinski *et al.* 1997) can estimate self-censorship of specific attitudes in survey responses, but would be an inefficient tool for studying willingness to express a wide variety of attitudes. More importantly, list experiments and other indirect techniques (like implicit association tests; e.g., Glaser and Knowles 2008) are specifically designed for measuring attitudes so sensitive that respondents might not admit to holding them *even in the privacy of a survey*—when they are uncomfortable confessing an attitude to an interviewer, an analyst, or even to themselves (Blair, Coppock, and Moor 2020). This is substantively distinct from my present goal of identifying how alternative social contexts induce different degrees and patterns of self-censorship (to test, e.g., whether heterodox views are suppressed in ideological enclaves).

Preference falsification, meanwhile, is difficult to define without reference to the individual's "true beliefs," which are hard to measure confidently when one suspects the individual of falsifying their preferences in the first place. Carlson and Settle (2016) cleverly contrived a social interaction in which study participants expressed their views in the form of public verbal responses to survey questions that they also answered in private. This political adaptation of the Asch (1955) paradigm allowed the authors to directly compare "public" and "private" opinions on the same scale, but it would be difficult to use a laboratory procedure like this to assess the prevalence of preference falsification at the population level. Moreover, it is impossible to measure self-censorship when participants are *required* to speak.

I propose a method that measures self-censorship and preference falsification in a relativistic framework, to characterize how different social contexts affect the content of political expression. Instead of querying respondents' attitudes about speech, I ask them whether they would use certain political words and phrases in certain social situations. I adapt document-scaling models to summarize their responses in two dimensions of "lexical ideology" and "outspokenness," and by implementing my measure in a survey experiment that varies the hypothetical social context of expression, I identify self-censorship and preference falsification as causal effects of social context on outspokenness and lexical ideology, respectively. My between-subjects design allows me to estimate self-censorship without invoking the politicized connotations of "freedom of speech," and avoids assumptions about respondents' "true beliefs" by operationalizing preference falsification as context-driven variation in word choices. This draws on the sociolinguistic phenomenon of *code-switching*: when language signifies social identity, speakers adopt different linguistic patterns in different social situations, according to the perceived norms of their audience (Wardhaugh 2015). In the sections that follow, I introduce the method I designed around this phenomenon, I validate it, and I demonstrate its application to the study of social influence over political expression.

2. Method

I employ a text ideal point model, of the type conventionally used to measure elites' ideological positions from convenience documents, but estimated on data derived from a novel survey instrument. My "What Would You Say?" question collects self-reported usage of political catchphrases, which I scale using Wordsticks—an ordinal modification of the Wordfish model. I find that by collecting data via a survey experiment, we can derive novel substantive applications for classic document-scaling methods.

2.1. Data Collection: The "What Would You Say?" Question

Figure 1a shows an example of the "What Would You Say?" (WWYS) question, which asks people whether they would use certain words and phrases, including slogans like BLACK LIVES MATTER, epithets like SNOWFLAKE, and constructs like TOXIC MASCULINITY and BIG GOVERNMENT. Political catchphrases such as these have attracted journalistic attention in recent years (e.g., Bui *et al.* 2022; Harmon 2021), and contemporary scholarship asserts that speakers' lexical choices have important consequences—in short, that WORDS MATTER (McConnell-Ginet 2020)—because they construct political debates around distinct frames, values, and assumptions. I use this lexical phenomenon to build a survey measure of political expression.

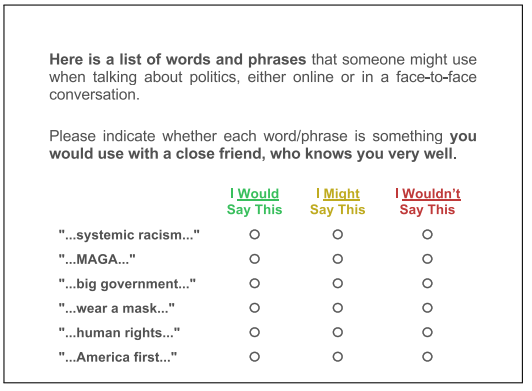
Political scientists have long exploited political lexica to scale documents. Laver and Garry (2000) were among the first to discover that individual words in party manifestos, such as TAXES and CHOICE, could furnish effective proxies for authors' ideologies. This insight, combined with the introduction of models like Wordscores (Laver, Benoit, and Garry 2003) and Wordfish (Slapin and Proksch 2008), provided an efficient way to summarize parties' policy positions. However, these methods are not typically used to study behaviors like self-censorship or preference falsification. This partly reflects limitations of observational analyses of pre-existing documents—limitations that the WWYS approach is designed to overcome.

First, the WWYS question avoids a major form of bias in convenience corpora, which only contain things that were said by people who chose to say them. This applies not only to elite manifestos but also to mass datasets such as tweets: most people avoid posting their political views online (Gallup 2022), and when studying self-censorship this amounts to selection on the dependent variable. By collecting data through a survey, the WWYS question facilitates question-driven corpus construction (Grimmer, Roberts, and Stewart 2022, ch. 3), allowing us to sample individuals who self-censor, and observe the things they *avoid* saying.

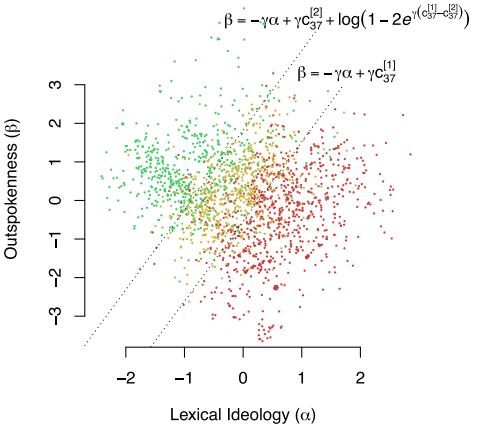
Second, the WWYS question makes it possible to manipulate (hypothetical) social context. The example in Figure 1a asks how respondents speak with "a close friend, who knows you very well," which theoretically elicits one's most authentic mode of self-presentation, but this can be replaced with alternative scenarios, including interactions with strangers, and in-group and out-group extremists, who may exert social influence. This makes it possible to estimate social contexts' effects on respondents' lexical choices in a relativistic framework, without conflating these effects with populist attitudes about freedom of speech. The WWYS question also facilitates best practices for causal inference with text outcomes: all covariates can be measured before treatment assignment, which is completely randomized (eliminating potential confounding due to self-selection into political conversations, see Settle and Carlson 2019), and the analyst conducts feature selection prior to data collection, limiting researcher degrees-of-freedom (Grimmer *et al.* 2022, ch. 24).

2.2. Model: Wordsticks

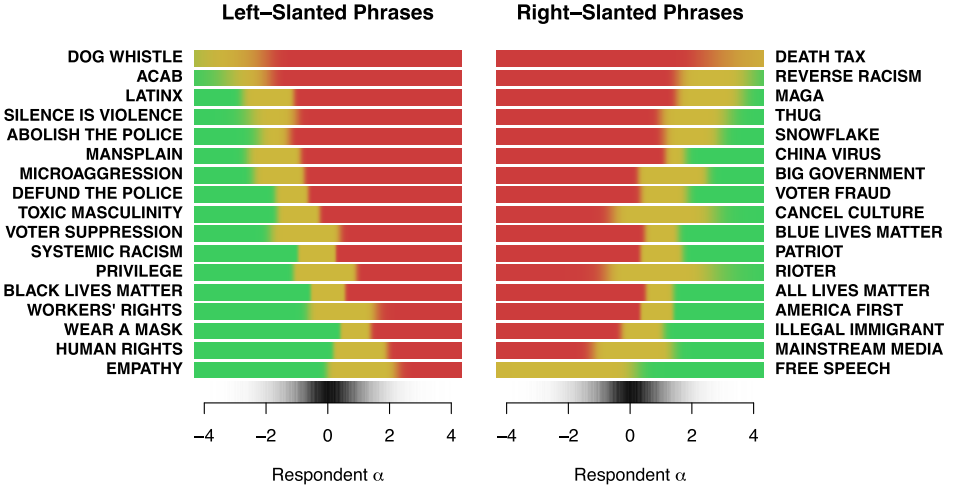
I also attenuate researcher-degrees-of-freedom by using an established modeling framework that focuses on my outcomes of interest. Despite its unique data-generating process, the responses derived from the WWYS question can be arranged into a matrix Y of dimension $I \times J$ that is structurally identical to the Document Term Matrix used in conventional text analyses. Rows i are people, columns j are phrases, and the only major difference is that the cells contain ordinal survey responses, instead of word



(a) Example "What Would You Say?" question.



(b) Responses to SYSTEMIC RACISM (color) by respondent lexical ideology (x axis) and outspokeness (y axis).



(c) Predicted phrase usage (color) as a function of respondent lexical ideology α (x axis).

Figure 1. The proposed method uses a novel survey question (a) to scale respondents (b) and phrases (c).

Panel (a) shows an example of the "What Would You Say?" (see Section 2.1) question with the three-point response scale used in Studies 1, 2, and 4. Study 3 used a four-point response scale. Typically 20 phrases are shown.

Panel (b) plots respondents' latent lexical ideology α (x-axis) and outspokenness β (y-axis) estimated from WWYS responses using the Wordsticks model (see Section 2.2), with points colored according to observed responses to the example phrase SYSTEMIC RACISM, and dashed lines marking boundaries between most-probable-response regions (see Equations (2) and (3)) estimated by Wordsticks. Data from Studies 1, 2, and 4, collected in 2021–22 ($N = 2,058$). Supplementary Appendix G provides similar plots for all phrases.

Panel (c) provides an ideological scaling of phrases, in terms of predicted response at different points of the lexical ideology (α) spectrum, for a respondent with $\beta = 0$. Thresholds between predicted response regions are estimated as c. Pixel values were averaged over 10,000 posterior draws, to create smooth gradients that represent estimation uncertainty. Rugplots on the x-axes similarly plot the (standard normal) posterior distribution of respondent lexical ideology α, to illustrate respondent density along this dimension. Supplementary Appendix H provides a version of this plot showing all phrases, as well as a plot of phrases from Study 3, which used a four-point scale. Supplementary Appendix F visualizes phrases' γ slant parameter estimates (which determine the direction of the gradient) in isolation. Data on phrases used in Studies 1, 2, and 4 (see Section 3) collected in 2021–22, based on Wordsticks scaling model (see Section 2) including pooled data from all studies, collected in 2021–23 (see Supplementary Appendix D).

counts. I therefore apply an item response theory (IRT) scaling model that is essentially an ordinal modification of Wordfish (Slapin and Proksch 2008), which I dub "Wordsticks":

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

where

$y_{ij} \in \{0,1,2\}$ the observed ordinal response to phrase j reported by respondent i,
$k \in \{1,2\}$ the possible response categories (excluding 0, the lowest category),
$\alpha_i \sim \mathcal{N}(0,1)$ respondent i's lexical ideology,
$\gamma_j \in \mathbb{R}$ phrase j's ideological slant,
$\beta_i \sim \mathcal{N}(0,\sigma_\beta^2)$ respondent i's outspokenness,
$c_j^{[k]} \in \mathbb{R}$ phrase j's response cutpoints.

Wordsticks estimates two respondent latent traits: α and β. The α trait corresponds to the left-right ideological position that is the object of interest in conventional document scaling studies; unlike prior studies, I do not interpret α as a proxy for some "true" ideology, and so I refer to α as "lexical ideology" to distinguish it from other operationalizations, but this parameterization is comparable to that of IRT models that have previously been used to estimate ideal points from, for example, legislators' roll-call votes (e.g. Clinton, Jackman, and Rivers 2004), judicial behavior (e.g. Martin and Quinn 2002), and mass survey responses (e.g. Treier and Hillygus 2009).

I refer to β as respondents' "outspokenness," which is a novel interpretation. In Wordfish, this is merely an intercept to absorb variation in manifestos' lengths (Slapin and Proksch 2008, 709). There is no such thing as "document length" in the WWYS question, since it presents the respondent with a fixed number of phrases, and asks whether the respondent would use *or avoid* each. So, in contrast to conventional analyses of documents, which only observe[1] the words an author *did* use, the WWYS question is able to detect abstention. Moreover, because it is a survey question, the WWYS method can sample individuals who vary in this trait to begin with, whereas conventional document scaling can only study authors of political texts, who are by definition outliers of extremely high political outspokenness.

Figure 1b plots respondents' positions in the two-dimensional latent space defined by lexical ideology α and outspokenness β (on the x- and y-axes, respectively—a convention retained in subsequent figures). Wordsticks estimates these respondent traits simultaneously with c and γ, which together define boundary lines

$$\beta = -\gamma_j\alpha + \gamma_j c_j^{[1]} \tag{2}$$

$$\text{and} \quad \beta = -\gamma_j\alpha + c_j^{[2]} + \log(1 - 2\exp(\gamma_j(c_j^{[1]} - c_j^{[2]}))) \tag{3}$$

that partition the α,β respondent space into regions where responses to phrase j are predicted to be "wouldn't," "might," and "would." Figure 1b plots these predicted response boundaries (dotted lines) along with observed responses (point color) for the example phrase SYSTEMIC RACISM. Because this phrase is left-slanted, respondents are more likely to say SYSTEMIC RACISM the further left is their lexical ideology α, and a person of fixed lexical ideology is more likely to say it the greater is their outspokenness β. This provides a spatial intuition for Wordsticks and illustrates the excellent fit between the model and the data for this phrase.

Figure 1c offers a more parsimonious ideological scaling of phrases, by plotting the expected response to each phrase (color) given the lexical ideology α (x axis) of a respondent with outspokenness $\beta = 0$. Shaded regions are delimited by the thresholds c, and smooth gradients were created by overlaying plots generated from all posterior samples of c and averaging their pixel values. This plot offers face validity

[1] When data are derived from counting terms occurring in natural documents, zero occurrences of a term represents *omission* but not necessarily deliberate abstention. Theoretically, we might say that omission approaches abstention as document length approaches infinity, but this is not a practical approach to measuring political expression.

(ex. ABOLISH THE POLICE scales to the left of DEFUND THE POLICE) and also has descriptive value in that it visualizes phrase extremity as distinct from slant (ex. ACAB is much more extreme than EMPATHY, though both are left-slanted). Moreover, the rightward slants of the phrases CANCEL CULTURE and FREE SPEECH validate one of my major motivations for developing the WWYS method in the first place: if the terminology by which we articulate issues of free expression is entangled with ideology, we cannot study these phenomena by directly querying respondents' opinions about these issues.

Supplementary Appendices C and D give further mathematical and estimation details, respectively, but in general, Wordsticks summarizes WWYS responses in two dimensions: a "left-right" lexical ideology dimension α, and what we can visualize as an "up-down" outspokenness dimension β. By implementing this method in a survey experiment that varies the social context of the WWYS question, and observing effects on lexical ideology and outspokenness, I derive a causal framework for estimating preference falsification/code-switching and self-censorship from survey data.

2.3. Limitations and Intended Use

The most obvious limitation of my method is that it does not directly observe natural speech in the wild. Laboratory and field experiments clearly remain the gold standard for observing real-world speech behaviors, and the WWYS method is not intended to replace these. Rather, the WWYS method is perhaps best understood as a *behavioral reframing* of the issue battery traditionally used to measure respondent ideology, composed of sentences that the respondent claims to "agree" or "disagree" with. This is arguably just a circuitous way of asking, "What Would You Say?" in the guise of asking "What Would You Agree With?" Indeed, public opinion surveys were originally developed to systematize the collection of public sentiments previously expressed through interpersonal conversation and other forms of real-world speech (Delli Carpini 2011).

However, scholars have long criticized traditional measures (Blumer 1948) for erasing the sociality of public opinion: when we express our opinions outside the artifice of a survey, we are almost always expressing ourselves *to somebody*. Traditional issue batteries also fail to capture respondents' intensity of feeling and commitment to advocate for their positions in everyday life (Ginsberg 1986, though see alternatives in Cavaille, Chen, and Van der Straeten Forthcoming). By remedying such shortcomings, the WWYS method contributes to survey methodology, and provides a means to explore contemporary issues surrounding social influence over political expression that are not accessible by direct, implicit, or trait-based methods. In this way, the WWYS approach complements existing tools for studying political expression.

3. Implementation

I implemented my method in five studies focused on the U.S. context, which are summarized in Table 1 (see Supplementary Appendix A for sample descriptives). I sought to recruit samples of U.S. adults evenly divided between self-described liberals, moderates, and conservatives, except in Study 4, which was conducted as part of a separate project (see Supplementary Appendix K.1) that targeted liberal and moderate Twitter users, and Study 5, which collected a representative sample of Facebook and Twitter users. Study 2 targeted Twitter and Facebook users, as a pilot for Study 5, which investigated social influence over online political expression. Social media is beyond the scope of the present discussion; I use the Study 2 data for its issue- and identity-based measures of ideology (see validation analyses below), and I use Study 5 solely as training data (see below), and not for any substantive analyses in this paper.

3.1. Social Context Treatments

I devised several context treatments to estimate effects on lexical ideology and outspokenness. Studies 1, 2, 3, and 5 all included the same "close friend" control condition, where (as shown in Figure 1a) respondents indicated "whether each word or phrase is something you would use *with a close friend,*

Table 1. Study information.

Study	N	Date	Panel	Social context treatment conditions
1	503	Spring 2021	Prolific	A close friend, who knows you very well Someone you just met, who doesn't know you too well
2	800	Fall 2021	MTurk	A close friend, who knows you very well Social media
3	901	Spring 2023	Prolific & MTurk	A close friend, who knows you very well The most liberal person you talk to (name) The most conservative person you talk to (name)
4	755	Summer 2022	MTurk	Twitter
5	2,339	Summer 2023	AmeriSpeak	A close friend, who knows you very well Twitter/Facebook

who knows you very well." This condition was chosen as a generic social context conducive to authentic self-presentation. To avoid assumptions about "true beliefs," I rely on the *relative* authenticity of self-presentation amongst close friends, and identify self-censorship and preference falsification as causal effects of other contexts *relative* to this substantive reference point (hence I describe the latter effects as *switches* rather than falsification).

Study 1 compared the "close friend" control to a treatment condition in which the respondent reported "whether each word or phrase is something you would use *with someone you just met, who doesn't know you too well*." This sought to evoke self-presentation to a new acquaintance, whose opinion of us is yet to be formed.

Study 3 sought to maximize ideological variation in interlocutors. Participants were asked to name the most liberal or most conservative person they talk to, and this name was piped into the WWYS question. These treatments sought to bookend the extent of left-right code-switching in participants' own conversation networks, by asking them how they talk to the most extreme people *that they actually talk to*. If individuals polarize their speech when talking to in-group members, or moderate their speech when talking to out-group members, it should be evident in their responses to these conditions.

This paper does not analyze treatment effects in Study 2, 4, or 5. In Study 4, only one condition was implemented: the WWYS question asked respondents to indicate "which words and phrases you would use *on Twitter*," to furnish data for validating the WWYS measure against "actual" speech in the form of respondents' tweets. Study 2 included various social media platforms as treatments, and Study 5 was a pre-registered replication of Study 2 (whose substance is the subject of a separate paper; I use it instrumentally here as a source of training data, as discussed in the next section).

3.2. Phrase Selection

The WWYS question affords the researcher great substantive control over the lexical space that defines their study, but this freedom must be exercised thoughtfully, to select a balanced set of phrases with relevance to one's inquiry—careless phrase selection risks introducing idiosyncratic researcher biases to the WWYS measure. Supplementary Appendix B gives a detailed account of my mixed-methods phrase selection procedure for these studies, but as a general principle, I suggest that a theory-driven selection of phrases should also be informed by a data-driven discovery exercise. I therefore applied a lasso-regularized binomial regression model to a set of 1 million tweets,[2] to identify terms most strongly predictive of users' ideology. This provided inspiration, external validity, and procedural reproducibility for my phrase selection.

[2] As noted in Section 2.1, many users self-censor on Twitter, however users self-censor in different ways, and I cast a wide net to capture a broad gamut of political expression to inform phrase selection.

However, this exercise also illustrated the limitations of observational text analysis for this project: terms that are predictive of user ideology are not necessarily ideological in a social or semantic sense (Supplementary Appendix B offers examples). I therefore did not limit myself to terms identified by the computational 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 study participants themselves (again, see Supplementary Appendix B).

Future researchers may adapt this procedure to define lexical spaces appropriate for their studies, but as a general rule, selected phrases should be face-valid to experts and familiar to participants, and analyses should be robust to estimation using various subsets of the terms. To this end, participants in Studies 1, 2, 3, and 5 responded to 20 phrases that were sampled from a larger superset, so that aggregate analyses marginalize over a diverse section of the American political lexicon (without overwhelming respondents). Studies 1 and 2 used a total set of 46 phrases, and Studies 3 and 5 used different (but overlapping) sets of 40 phrases, and the first six phrases shown were held constant. For Study 4, a briefer version of the WWYS question was composed from 11 fixed items (see Supplementary Appendix K.2).

To avoid SUTVA violations that can arise in causal inference with latent variables (Egami *et al.* 2022), I estimated Wordsticks first on data from Studies 2, 4, and 5, extracted the phrase parameter estimates, and passed them to a second-stage model (see Supplementary Appendix D.5) that held these parameters fixed while estimating α and β for the respondents in Studies 1 and 3 (which contained the experiments I analyze in Section 4.2).

4. Results

I present results with two goals: multi-pronged descriptive validation of the α and β estimates derived from the WWYS method, and demonstration of the types of experimental analyses this method can afford. All non-binary variables are standardized.

4.1. Validations and Descriptive Analyses

I validate the WWYS measure in several different ways. First, I verify the ideological nature of the α trait, lexical ideology, and I explore its relationship to issue-based and identity-based operationalizations of ideology. Figure 2a shows that lexical ideology tracks closely with issue ideology, constructed by Principal Components Analysis of an issue preference battery (see Supplementary Appendix I). Figure 2b illustrates that α also correlates with respondents' identification with the terms "liberal" and "conservative," constructed by Principal Components Analysis of standard measures of identity importance, descriptiveness, and "we/they" identification (Huddy, Mason, and Aarøe 2015; see Appendix J).

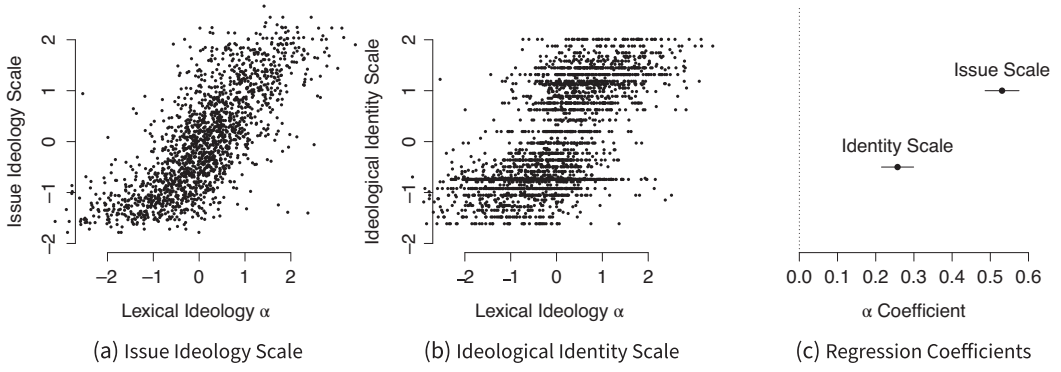


Figure 2. Validation of lexical ideology estimates against (a) issue ideology derived from one-dimensional scaling of 10 issue preference items, and (b) signed identity strength constructed by one-dimensional scaling of three ideological identity questions and interacting with a signed indicator for liberal/conservative identification. Panel (c) plots regression coefficients (with 95% CIs) for both of these measures predicting α (see Table 2, model 4). Data from Studies 2 and 3, collected in 2021 and 2023 (N = 1,701).

Figure 2c confirms that both of these operationalizations of ideology are independently predictive of α in a linear regression (see Table 2, model 4).

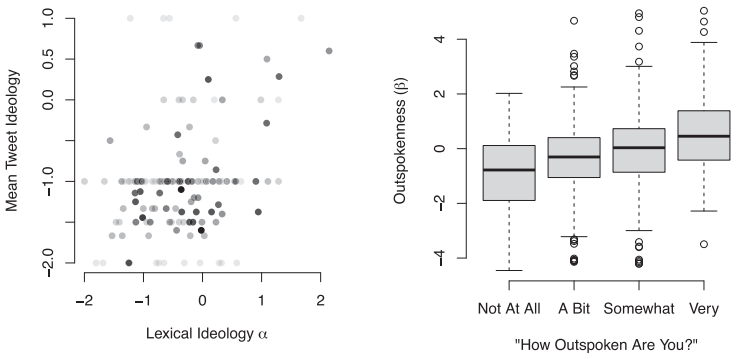
Next, I validate that the WWYS measure correlates with corresponding summaries of behavioral speech, using tweets collected from participants in Study 4. Study 4 was an experiment that recruited left-of-center Twitter users and collected their tweets for several weeks. The substance of this experiment is beyond the scope of the present paper; it is relevant solely because it included a brief version of the WWYS question and also collected respondents' tweets, and so permits a behavioral validation of the α and β traits. Study 4 was not designed for this purpose, though, so it has some shortcomings: the estimation of α and β is less precise than in the other studies due to the use of a brief version of the WWYS question; there are fewer tweets-per-user than would be ideal for this validation exercise because data collection was focused on a narrow time period; and there is not as much ideological variation as would be ideal because recruitment targeted left-of-center users. Nonetheless, these shortcomings are tolerable in the interest of conducting a behavioral validation of the WWYS measure.

I compared respondents' latent trait estimates to hand labels (Lowe and Benoit 2013) measuring ideological content of their tweets. A total of 2,000 tweets from 265 users were manually reviewed by a team of three coders, who labeled each tweet as "very liberal," "liberal," or "conservative," and also reported whether they felt "sure," "not so sure," or had "no idea at all" about this ideology label (see Supplementary Appendix K for details).

To validate the lexical ideology estimates, I took the mean of the ideology labels of each user's tweets. This entailed dropping all "no idea at all" tweets, and focusing on the 561 remaining tweets with ideology labels. Mean tweet ideology was taken by coding "very liberal" tweets as −2, "liberal" tweets as −1, and "conservative" tweets as +1, and then taking the mean over all such tweets for each user. This produced aggregated ideology estimates for 172 users (dropping 93 users with no labeled ideological tweets). I then regressed these hand-labeled tweet ideology estimates on the WWYS lexical ideology estimates, and found a strong linear relationship ($p < 0.01$; Pearson's $\rho = 0.29$, $p < 0.001$; see Supplementary

Table 2. Alpha regressions.

	Treatments (Pooled)	Treatments (Test Only)	All Ideo Measures	Issues and Identity
	(1)	(2)	(3)	(4)
t = Stranger	− 0.04 (0.06)	− 0.04 (0.06)		
t = Liberal	− 0.22 (0.06)	− 0.21 (0.06)	− 0.17 (0.05)	− 0.17 (0.05)
t = Conservative	0.21 (0.06)	0.21 (0.06)	0.21 (0.05)	0.21 (0.05)
t = Social media	0.09 (0.05)		0.10 (0.05)	0.10 (0.05)
Issue scale			0.51 (0.02)	0.53 (0.02)
Identity scale			0.14 (0.04)	0.26 (0.02)
Age (decades)	0.10 (0.01)	0.09 (0.02)	0.08 (0.01)	0.09 (0.01)
6-point ideo	0.48 (0.03)	0.50 (0.03)	0.15 (0.04)	
6-point PID	0.22 (0.03)	0.19 (0.03)		
College	− 0.07 (0.03)	− 0.03 (0.04)	− 0.04 (0.03)	− 0.04 (0.03)
Male	0.20 (0.03)	0.21 (0.04)	0.11 (0.03)	0.10 (0.03)
POC	− 0.04 (0.04)	− 0.05 (0.05)	− 0.02 (0.04)	− 0.02 (0.04)
Study 2 FE	0.09 (0.05)			
Study 3 FE	− 0.03 (0.06)	− 0.03 (0.06)	− 0.15 (0.04)	− 0.15 (0.04)
Intercept	− 0.55 (0.07)	− 0.56 (0.08)	− 0.28 (0.07)	− 0.28 (0.07)
Observations	2,200	1,401	1,701	1,701
Adjusted R^2	0.54	0.51	0.63	0.62



(a) Tweet validation of α trait. (b) Self-report validation of β trait.

Figure 3. Additional validations of α and β estimates: Panel (a) plots lexical ideology (α) against ideology scores from hand-labeled tweets ($p < 0.01$, R^2=.08, data collected in Study 4, 2022, $N = 148$). Points represent users, and are shaded according to the number of tweets available from that user, ranging from 1 tweet (light grey) to 10 tweets (black). Panel (b) plots outspokenness (β) against participants' responses to the question, "Generally speaking, how outspoken are you about politics and current events?" ($p < 0.001$, R^2=.09, data collected in Study 3, 2023, $N = 901$).

Appendix K). This analysis is limited by the roughness of the tweet ideology metric, due to the small size of the dataset: the modal number of tweets-per-user was 2, and many users had only one ideological tweet available for analysis. This validation is also made quite conservative by the left-leaning composition of the survey sample, which limits the extent of ideological variation we can analyze. Nonetheless, Figure 3a, which plots the relationship between lexical ideology and mean tweet ideology (shading points according to the number of tweets available), displays a visible relationship between the lexical ideology trait estimated by the WWYS method and the human-rated ideology of users' online speech.

I also tested whether the WWYS outspokenness estimates corresponded to the hand-labeled "sureness" that annotators reported regarding the ideological coding of users' tweets. I included all 2,000 tweets from 265 users (including the 93 users with no labeled ideological tweets), and took the user-level mean over the sureness labels, where "no idea at all" was coded as 0, "not so sure" was coded as 1, and "sure" was coded as 2. I then regressed hand-labeled sureness on the WWYS outspokenness estimates. The relationship meets conventional standards of statistical significance in a linear regression ($p < 0.05$; Pearson's $\rho = 0.16$, $p < 0.05$; see Supplementary Appendix K), although it is noisier than in the ideology analysis. This was not the intended purpose of collecting sureness ratings, and the analysis is again limited by the size of the dataset: because political tweets are rare, a small sample may fail to include any even if the user does sometimes post about politics, such that this dataset fails to capture variation at the lower end of political tweeting. Nonetheless, this analysis helps validate the β trait as "outspokenness," since the more politically outspoken an individual is, we should expect them to express political opinions more frequently and discernibly.

Separate from the above behavioral validation of the outspokenness estimates, I sought to validate my *labelling* of β as "outspokenness." So, I tested whether β correlated with respondents' self-rated outspokenness, measured with the question, "Generally speaking, how outspoken are you about politics and current events?" which was included in Study 3. Figure 3b illustrates the positive relationship between these responses and the latent β estimates (Pearson's $\rho = 0.31$, $p <$ 1e-)20. As discussed in the introduction, a direct question like this lacks substantive interpretability, but it is useful for validating the labeling of β as "outspokenness," since it correlates with respondents' self-labeling as such.

My final descriptive analyses employ linear regression to characterize political and demographic predictors of the α and β traits. Figure 4a shows that lexical ideology α correlates strongly with conventional Likert-scale measures of liberal-conservative ideology and partisanship, and with several demographic variables: college education is associated with somewhat more left-leaning speech, while both age and male gender are associated with more right-leaning speech. This is consistent with past

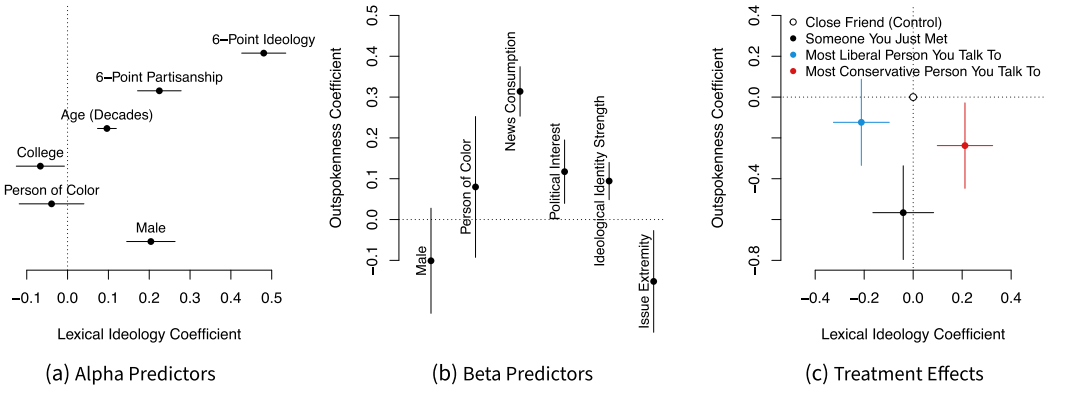


Figure 4. Linear regression coefficients (and 95% CIs) for descriptive predictors and context treatment effects. Panel (a) plots the descriptive predictors of lexical ideology α (see Table 2, model 1, $N = 2,200$; data from Studies 1, 2, and 3, collected in 2021–23). Panel (b) plots the descriptive predictors of outspokenness β (see Table 3, model 1, $N = 1,698$; data from Studies 2 and 3, collected in 2021–23). Panel (c) plots the effects of social context treatments: the effects of the "someone you just met" treatment (black) and the "most liberal/conservative person you talk to" treatments (blue/red) are plotted relative to the "close friend" control (black circle, at origin by construction). In this two-dimensional plot, the *x*-axis denotes left-right shift in lexical ideology α (see Table 2, model 2, $N = 1,401$), and the *y*-axis denotes up-down shifts in outspokenness β (see Table 3, model 3, $N = 1,401$; Data from Studies 1 and 3, collected in 2021–23).

evidence of gender differences in political speech (Karpowitz and Mendelberg 2014; Monroe, Colaresi, and Quinn 2008; Schwartz *et al.* 2013, c.f. Gilligan 1993); the age finding, meanwhile, may reflect that the Left often replaces "old-fashioned" terminologies with more progressive coinages, which may be unfamiliar to older people.

Figure 4b offers a similar descriptive analysis of outspokenness β, which is significantly predicted by self-reported news consumption (Supplementary Appendix L disaggregates this by news sources), consistent with existing evidence that news consumption predicts engaging in political discussion (Hao, Wen, and George 2014; Mcleod, Scheufele, and Moy 1999). It is also associated with political interest (whose estimate becomes larger when news is excluded; again see Supplementary Appendix L), which is a sensible finding that accords with Barberá's (2015) interpretation of this parameter in the Tweetscores model of ideology, and which is also consistent with past evidence on political engagement generally (e.g., Eveland and Hively 2009; Verba, Lehman Schlozman, and Brady 1995). Ideological identity strength is also a predictor of outspokenness, but issue extremity (the absolute value of issue ideology) takes a negative-signed coefficient, similar to findings from Mason (2018) that identity strength is a better predictor of political activism than issue position.

These analyses offer a multiplex validation of the WWYS measure. Though it does not directly measure speech, it correlates with the best available behavioral measures, it accords with respondents' self-identification in terms of ideology and outspokenness, and it has descriptive properties consistent with other studies of political expression.

4.2. Experimental Analyses

I now analyze the experiments in Studies 1 and 3. Figure 4c plots the social context treatment effects in both dimensions simultaneously: left-right shifts in lexical ideology α are plotted on the *x*-axis, and up-down shifts in outspokenness β are plotted on the *y*-axis. I estimate the effects of three different social context treatments by comparing them to speaking with "a close friend, who knows you very well," as a common control. I find that the treatment (in Study 1) of speaking with "someone you just met, who doesn't know you too well" (black) had a null effect on participants' lexical ideal point α, but it induced a large *downward* shift in outspokenness, with magnitude greater than any descriptive predictor of outspokenness. Substantively, individuals appear to self-censor ideological terms when speaking with strangers, as opposed to close friends. Meanwhile, I find that the treatments (in Study 3) of speaking with

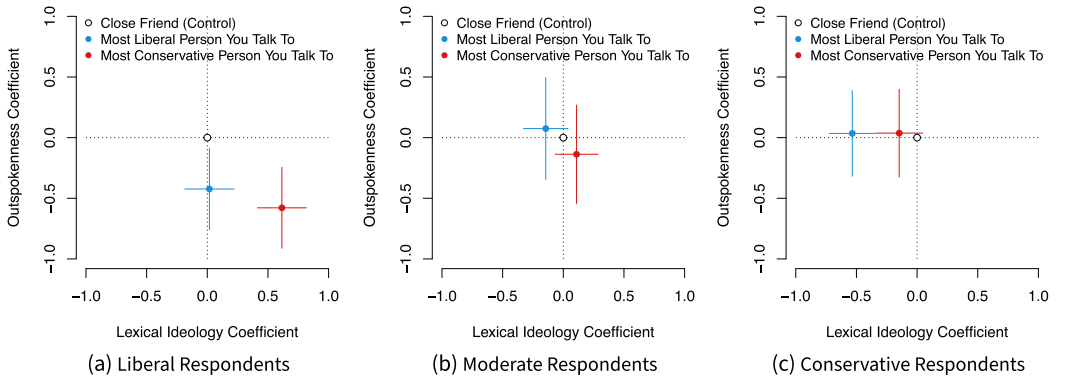


Figure 5. Study 3 treatment coefficients (and 95% CIs) by respondent ideology: Liberals (Panel (a); see model 1 of Supplementary Tables 7 and 8), Moderates (Panel (b); see model 2 of Supplementary Tables 7 and 8), and Conservatives (Panel (c); see model 3 of Supplementary Tables 7 and 8). Data from Studies 1 and 3, collected in 2021–23 (see Supplementary Appendix M).

the most liberal (blue) and most conservative (red) person one talks to, induced leftward and rightward shifts in lexical ideology (respectively).

However, when I disaggregate the Study 3 treatment effects by respondent ideology (Figure 5), code-switching appears to be driven primarily by liberals and conservatives *accommodating* their out-groups. This contradicts Social Identity Theory (Turner 1991), which predicts that individuals conform to in-groups more than out-groups, and that their attempts to align themselves with the prototypical positions of their in-groups result in polarization (Abrams *et al.* 1990; McGarty *et al.* 1992). On the contrary, I find that *out-group* interlocutors induce significant *depolarization*, while both the in-group treatment effects are insignificant at conventional levels and have signs opposite to that predicted by polarization theories. Moderates, meanwhile, exhibit left-right code-switching effects that are not distinct from the null at conventional levels, though posterior power analyses (see Supplementary Appendix M) indicate that Study 3 was under-powered to detect effects of this magnitude, and so the lack of statistical significance should not be interpreted as evidence of null effects.

To further characterize these treatment effects, Figure 6 plots the phrase-level effects of the "most liberal person" and "most conservative person" treatments, subset by the self-described ideology of the respondent (coarsened to liberal/conservative, including leaners). As expected, the "most liberal" treatment tends to increase usage of liberal terms and decrease usage of conservative terms, while the "most conservative" treatment has the opposite effect. Consistent with depolarization, the outgroup treatments affect more phrases than ingroup treatments, although there is some evidence of phrase-level polarization: liberal respondents with liberal interlocutors are more likely to say EAT THE RICH, and less likely to say REVERSE RACISM and MAINSTREAM MEDIA; conservative respondents with conservative interlocutors are more likely to say SNOWFLAKE and MAINSTREAM MEDIA, and less likely to say ABORTION IS HEALTHCARE, HUMAN RIGHTS, and HETERONORMATIVE.

This adds some nuance to the analyses presented above, although the interpretation of the phrase-level polarization evidence depends on how one construes these specific phrases substantively (discussed further below). This illustrates an advantage of the WWYS method relative to direct questions, list experiments, and implicit association tests: it not only furnishes overall quantitative summaries of expression, but also provides richer insights into its content, amenable to substantive interpretation.

5. Discussion

Overall, my demonstration experiments depict Americans not as mercenary preference-falsifiers who polarize discourse to appease their in-groups, but as meek self-censorers who muffle their extremity in interactions with out-groups. There is also no indication that conservatives are more prone to self-censorship than liberals (if anything, Figure 5 suggests the opposite). This illustrates how the WWYS method can speak to contemporary debates about political expression.

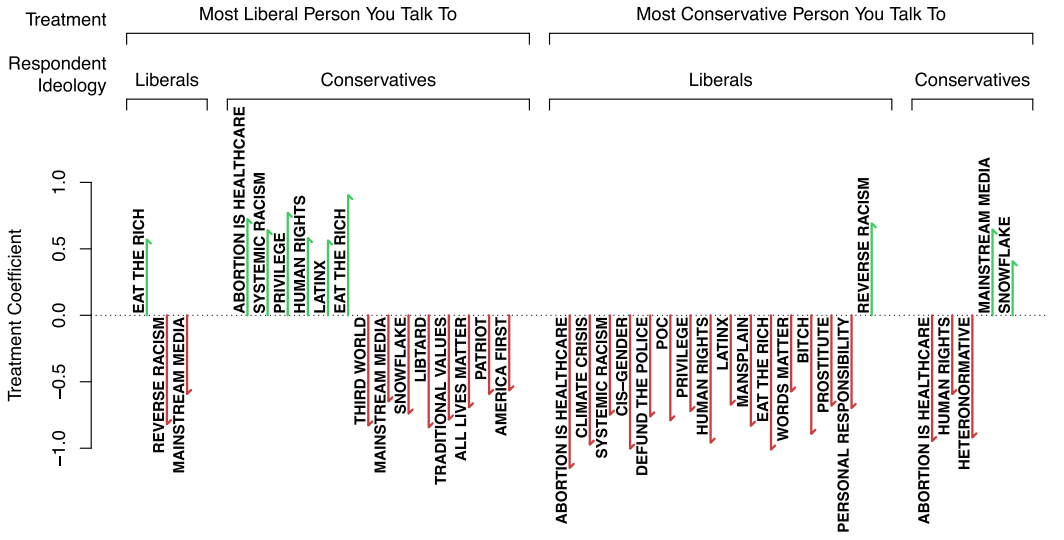


Figure 6. Phrase-level effects of "most liberal/conservative person you know" treatments, subset by treatment condition and respondent ideology (see annotations at top). Barbs plotted in green (red) point upward (downward) in proportion to the increase (decrease) in phrase usage induced by the treatment in each subset of respondents. Ordinal logit models were estimated for all 40 phrases included in Study 3. False discovery was attenuated by Benjamini–Hochberg correction; phrases are only plotted if $p_{\text{corrected}} < 0.05$. Within each subgroup, phrases are ordered horizontally by `rank(gamma)` so that left-slanted phrases appear to the left of right-slanted phrases. Data from Study 3, collected in 2023 ($N = 901$).

These results might be influenced by homophily—humans' tendency to associate with similar others (McPherson, Smith-Lovin, and Cook 2001). When asked to think of a close friend, liberals will think of liberals, and conservatives will think of conservatives, which could attenuate the in-group treatment effects (and enlarge the out-group treatment effects), if the close-friend control has effectively moved closer to the in-group treatment (and further from the out-group treatment). However, this attenuation is undone if the whole ideological distribution of one's interlocutors shifts with one's own ideology. Although ceiling effects could attenuate in-group treatment effects among the most extreme respondents (whose close friends are so extreme that there is nobody more extreme in their network), these respondents are by definition a minority. Moreover, the treatments were intentionally designed to reference the respondent's actual discussion network, in order to accurately describe the range of ideological code-switching in the social interactions people actually have, so even if in-group effects are attenuated, this attenuation describes a substantive reality.

Another consideration is that, if respondents engage in social self-censorship, they may also give dishonest survey responses. However, the WWYS method's between-subjects design causes dishonesty to manifest simply as an attenuation of treatment effects. Suppose a respondent considers a phrase so socially (un)desirable that they always report (not) using it. This would distort their point estimates of α and β, but it only affects causal estimates by attenuating the range of possible responses that could be induced by different treatments. So long as respondents' perceptions of phrases' social desirability (as a survey response) do not depend on the context in which the phrase is used (in the WWYS question), the only distortion to the treatment effects will be attenuation.

Readers might be concerned that the code-switching effects in Study 3 do not indicate shifts in respondents' position-taking, but rather respondents' anticipated need to reference concepts raised by their interlocutor. However, Study 3 included the following instruction (appended to the prompt displayed in Figure 1a):

> Note: Please only consider whether you would use a phrase sincerely. It doesn't count if you would only use a phrase sarcastically, or only to quote someone else who said it, or only as a joke.

This clarifies the interpretation of the question and resulting data, and explicitly rules out a referential interpretation[3] of the Study 3 code-switching effects.

More broadly, though, one might still question whether these effects pertain more to the style or the substance of speech. Empirically, the WWYS question appears to capture a combination of issue positions and identity signals, and the phrase-level treatment effects offer examples consistent with both a style and a substance interpretation. For example, conservatives' avoidance of the word LIBTARD when talking to a liberal may be simple politeness. Similarly, a person could make an argument for economic redistribution without saying EAT THE RICH, or for a unilateral approach to foreign policy without saying AMERICA FIRST. However, it is hard to divorce broad shifts in the lexical ideology of speech from its positional content. When liberals avoid saying ABORTION IS HEALTHCARE to conservatives, they forsake a framing that is inherently positional, and when conservatives adopt the phrase SYSTEMIC RACISM in conversations with liberals, they concede the substantive point that racism is in fact a systemic phenomenon. Perhaps the real question is whether these concessions are *strategic*: do speakers adopt the language of their out-group in order to make their overall argument more persuasive? Do they do it to avoid conflict? In either case, are these lexical strategies actually effective? These questions are beyond the scope of the present research, but would be excellent topics for future studies.

Table 3. Beta regressions.

	With Issues	Without Issues	Treatments (Test Only)
	(1)	(2)	(3)
Age (decades)	0.03 (0.03)	−0.01 (0.02)	0.001 (0.03)
6-point ideo	−0.01 (0.06)	−0.07 (0.04)	0.02 (0.06)
6-point PID	−0.05 (0.06)	−0.02 (0.04)	−0.01 (0.06)
Political interest	0.12 (0.04)	0.18 (0.03)	0.05 (0.04)
Identity strength	0.09 (0.02)	0.08 (0.02)	0.12 (0.03)
POC	0.08 (0.09)	0.13 (0.07)	0.11 (0.10)
Male	−0.10 (0.07)	−0.18 (0.05)	−0.20 (0.07)
News scale	0.31 (0.03)	0.30 (0.03)	0.33 (0.03)
t = Stranger		−0.57 (0.11)	−0.57 (0.12)
t = Liberal	−0.10 (0.11)	−0.10 (0.10)	−0.12 (0.11)
t = Conservative	−0.23 (0.11)	−0.22 (0.10)	−0.24 (0.11)
t = Social media	−0.24 (0.10)	−0.27 (0.10)	
abs(Issue scale)	−0.15 (0.06)		
Study 2 FE		−0.31 (0.09)	
Study 3 FE	−0.29 (0.09)	−0.60 (0.11)	−0.59 (0.11)
Study 4 FE		−0.42 (0.14)	
Intercept	0.17 (0.14)	0.54 (0.12)	0.49 (0.16)
Observations	1,698	2,954	1,401
Adjusted R^2	0.15	0.14	0.16

[3]A referential interpretation is also inconsistent with these effects manifesting mostly through *avoidance* of phrases.

6. Conclusion

To be sure, political speech is not reducible to the language a speaker uses to make their point. However, by crafting a survey question around the concrete act of using or avoiding certain political lexica, we can summarize how a survey respondent would self-reportedly present themselves in different social situations. This approach is more feasible[4] than directly observing private conversations, it is less sensitive to the political connotations of "free speech" that make direct questions more attitudinal than behavioral, and it provides much more insight into the content of speech than is available from indirect methods like list experiments or implicit association tests. In essence, the method introduced in this paper adapts a set of tools originally developed for scaling elite documents, and applies them to an improved survey instrument for characterizing social influence over mass political expression.

By implementing my measure in a survey experiment, and randomizing the hypothetical speech context of the WWYS question, I operationalize self-censorship and preference falsification as causal treatment effects on outspokenness and lexical ideology. Relative to how they speak with close friends, I find that individuals self-censor political catchphrases when speaking to new acquaintances, and conform to their most liberal and conservative interlocutors by shifting their lexical ideology leftward and rightward, respectively. If we accept the idea of close friendship as a reference point for authentic self-presentation, my results indicate that preference falsification favors moderation, rather than polarization, which contradicts popular claims that political discourse is polarized by ideological conformity.

However, this concept of authenticity merits critique. The very notion of preference falsification implies that expressed preferences can be described as "true" or "false"—a binary I have sought to avoid in this project, by adopting a framework in which I assume people are most authentic with their close friends, and define self-censorship and preference falsification in relation to this reference point. If we were to adopt a fully relativistic framework, we might reject the idea that any one speech setting is more authentic than another. If all social behavior is performative (e.g., Butler 1999; Goffman 1956), then we might treat expression in each context as distinct but equally valid performances; this is why I advocate describing the differences between these performances as code-switching, rather than preference falsification.

Arguably, what matters most is not whether speech is authentic or inauthentic, but what conditions tend to polarize or depolarize it. In this framing, we could conclude that ideological homophily—the tendency of liberals to associate with liberals, and conservatives with conservatives—threatens to polarize political discourse, by reducing the frequency of cross-ideological encounters that moderate individuals' speech (according to Study 3). In this way, my findings resonate with the prejudice-reduction effects of inter-group contact (e.g., Kalla and Broockman 2020).

My tests were not pre-registered, and the samples I analyze are drawn from non-representative online panels, so the results should not be considered conclusive. However, they demonstrate the kinds of analyses that the WWYS method makes possible. My aim is not to close debates about social influence over political discourse, but to advance these debates by supplying a tool to collect population-level descriptive and causal data on speech behaviors. Though my findings cast doubt on claims that discourse is polarized by conformity to ideological in-groups, it remains possible that polarizing conformity is more prevalent online, due to the unique affordances of social media platforms (Sunstein 2017). The latter possibility is the subject of a separate study that also uses the WWYS measure, to estimate differences between online and offline speech.

Future studies can probe other open questions raised by this paper. For example, the relationship I observe between outspokenness and news consumption is intriguing, particularly since the most "verbal" news sources—discussions, radio, and podcasts—take the largest coefficients (see Supplementary Appendix L). This suggests a potential mechanism by which political catchphrases are learned, or perhaps normalized. More broadly, it would be interesting to learn more about the process by which catchphrases are coined and adopted: this could be achieved by embedding the WWYS question in a

[4]Not to mention ethical (cf. Henle and Hubbell 1938).

panel survey, and analyzing responses longitudinally alongside contemporaneous data on news media content and consumption. Further research can also extend the WWYS method to non-U.S. settings, particularly by incorporating a multidimensional ideology space into Wordsticks (I have focused on the United States because it is my area of substantive expertise, but I expect that the existence of political lexica is quite a general phenomenon). Ultimately, I hope that this method will help advance our understanding of social influence over individual political expression, and its implications for ordinary citizens' collective discourse.

Acknowledgments. I gratefully acknowledge input on prior drafts of this paper from Andy Guess, Tali Mendelberg, Markus Prior, Brandon Stewart, Pablo Barberá, Danny Daneri, Kevin Munger, Will Lowe, Dean Knox, Lisa Argyle, and participants at PolComm 2021 and in the American Political Behavior Research Group at Princeton University. All remaining errors are my own. I acknowledge using ChatGPT (without modification, in June 2023) to enhance readability in Supplementary Appendix B figures, and to write a loop for multiple-testing correction for Figure 6.

Funding. This work was supported by the Center for the Study of Democratic Politics at Princeton University, and by the TESS Special Competition Using Targeted Samples.

Competing Interest. The author declares no competing interests.

Research with Human Subjects. This research was approved by Princeton University IRB #12687, #13356, #13544, #13894, and #15208.

Data Availability Statement. Replication code for this article has been published in Code Ocean, a computational reproducibility platform that enables users to run the code and can be viewed interactively at https://codeocean.com/capsule/9261446/tree/v3. A preservation copy of the same code and data can also be accessed via Dataverse at https://doi.org/10.7910/DVN/GBSOAS (Schulz 2024).

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2024.15.

References

Abrams, D., M. Wetherell, S. Cochrane, M. A. Hogg, and J. C. Turner. 1990. "Knowing What to Think by Knowing Who You Are." *British Journal of Social Psychology* 29 (2): 97–119.

Ackerman, E., et al. 2020. "A Letter on Justice and Open Debate." https://harpers.org/a-letter-on-justice-and-open-debate/.

Asch, S. E. 1955. "Opinions and Social Pressure." *Scientific American* 193 (5): 31–35.

Barberá, P. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23 (1): 76–91.

Blair, G., A. Coppock, and M. Moor. 2020. "When to Worry about Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments." *American Political Science Review* 114 (4): 1297–1315.

Blair, G., and K. Imai. 2012. "Statistical Analysis of List Experiments." *Political Analysis* 20 (1): 47–77.

Blumer, H. 1948. "Public Opinion and Public Opinion Polling." *American Sociological Review* 13 (5): 542–549.

Bui, Q., S. Chodosh, J. Bennett, and J. McWhorter. 2022. "You Can't Say That!" *The New York Times*, December 22.

Butler, J. 1999. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge.

Carlson, T. N., and J. E. Settle. 2016. "Political Chameleons: An Exploration of Conformity in Political Discussions." *Political Behavior* 38 (4): 817–859.

Cavaille, C., D. L. Chen, and K. Van der Straeten. Forthcoming. "Measuring Differences in Preference Intensity." *Political Science Research and Methods*.

Chong, D., J. Citrin, and M. Levy. 2022. "The Realignment of Political Tolerance in the United States." *Perspectives on Politics* 22 (1): 131–152.

Clinton, J., S. Jackman, and D. Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98 (2): 355–370.

Delli Carpini, M. X. 2011. "Constructing Public Opinion: A Brief History of Survey Research." In *The Oxford Handbook of American Public Opinion and the Media*, 284–301. Oxford: Oxford University Press.

Egami, N., C. J. Fong, J. Grimmer, M. E. Roberts, and B. M. Stewart. 2022. "How to Make Causal Inferences Using Texts." *Science Advances* 8 (42): eabg2652.

Eveland, W. P., and M. H. Hively. 2009. "Political Discussion Frequency, Network Size, and 'Heterogeneity' of Discussion as Predictors of Political Knowledge and Participation." *Journal of Communication* 59 (2): 205–224.

Gallup. 2022. "Media and Democracy: Unpacking America's Complex Views on the Digital Public Square." Technical report, Knight Foundation.

Gibson, J. L., and J. L. Sutherland. 2023. "Keeping Your Mouth Shut: Spiraling Self-Censorship in the United States." *Political Studies Quarterly* 138 (3): 361–376.
Gilligan, C. 1993. *In a Different Voice: Psychological Theory and Women's Development*. Cambridge: Harvard University Press.
Ginsberg, B. 1986. *The Captive Public: How Mass Opinion Promotes State Power*. New York: Basic Books.
Glaser, J., and E. D. Knowles. 2008. "Implicit Motivation to Control Prejudice." *Journal of Experimental Social Psychology* 44 (1): 164–172.
Goffman, E. 1956. *The Presentation of Self in Everyday Life*. University of Edinburgh Social Sciences Research Centre Monographs. Edinburgh: University of Edinburgh Press.
Grimmer, J., M. E. Roberts, and B. M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press.
Habermas, J. 1989. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Translated by T. Burger. Cambridge: The MIT Press.
Hao, X., N. Wen, and C. George. 2014. "News Consumption and Political and Civic Engagement among Young People." *Journal of Youth Studies* 17 (9): 1221–1238.
Harmon, A. 2021. "BIPOC or POC? Equity or Equality? The Debate over Language on the Left." *The New York Times*, November 1.
Hayes, A. F., B. R. Uldall, and C. J. Glynn. 2010. "Validating the Willingness to Self-Censor Scale II: Inhibition of Opinion Expression in a Conversational Setting." *Communication Methods and Measures* 4 (3): 256–272.
Henle, M., and M. B. Hubbell. 1938. "Egocentricity in Adult Conversation." *The Journal of Social Psychology* 9 (2): 227–234.
Huddy, L., L. Mason, and L. Aarøe. 2015. "Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity." *American Political Science Review* 109 (1): 1–17.
Kalla, J. L., and D. E. Broockman. 2020. "Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments." *American Political Science Review* 114 (2): 410–425.
Karpowitz, C., and T. Mendelberg. 2014. *The Silent Sex: Gender, Deliberation, and Institutions*. Princeton: Princeton University Press.
Kuklinski, J. H., P. M. Sniderman, K. Knight, T. Piazza, P. E. G. R., and B. Mellers. 1997. "Racial Prejudice and Attitudes toward Affirmative Action." *American Journal of Political Science* 41 (2): 402–419.
Kuran, T. 1995. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge: Harvard University Press.
Laver, M., K. Benoit, and J. Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97 (2): 311–331.
Laver, M., and J. Garry. 2000. "Estimating Policy Positions from Political Texts." *American Journal of Political Science* 44 (3): 619–634.
Lowe, W., and K. Benoit. 2013. "Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark." *Political Analysis* 21 (3): 298–313.
Martin, A. D., and K. M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10 (2): 134–153.
Mason, L. 2018. *Uncivil Agreement: How Politics Became our Identity*. Chicago: University of Chicago Press.
McConnell-Ginet, S. 2020. *Words Matter: Meaning and Power*. Cambridge: Cambridge University Press.
McGarty, C., J. C. Turner, M. A. Hogg, B. David, and M. S. Wetherell. 1992. "Group Polarization as Conformity to the Prototypical Group Member." *British Journal of Social Psychology* 31: 1–20.
Mcleod, J. M., D. A. Scheufele, and P. Moy. 1999. "Community, Communication, and Participation: The Role of Mass Media and Interpersonal Discussion in Local Political Participation." *Political Communication* 16 (3): 315–336.
McPherson, M., L. Smith-Lovin, and J. M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27 (1): 415–444.
Menzner, J., and R. Traunmüller. 2023. "Subjective Freedom of Speech: Why Do Citizens Think They Cannot Speak Freely?" *Politische Vierteljahresschrift* 64 (1): 155–181.
Monroe, B. L., M. P. Colaresi, and K. M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16 (4): 372–403.
Noelle-Neumann, E. 1974. "The Spiral of Silence a Theory of Public Opinion." *Journal of Communication* 24 (2): 43–51.
Norris, P. 2023. "Cancel Culture: Myth or Reality?" *Political Studies* 71 (1): 145–174.
Revers, M., and R. Traunmüller. 2020. "Is Free Speech in Danger on University Campus? Some Preliminary Evidence from a Most Likely Case." *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 72 (3): 471–497.
Schwartz, H. A., et al. 2013. "Personality, Gender, and Age in the Language of Social Media." *PLoS One* 8 (9): e73791.
Settle, J. E., and T. N. Carlson. 2019. "Opting Out of Political Discussions." *Political Communication* 36 (3): 476–496.
Slapin, J. B., and S.-O. Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–722.
Stouffer, S. A. 1955. *Communism, Conformity, and Civil Liberties: A Cross-Section of the Nation Speaks its Mind*. Garden City: Doubleday.

Sunstein, C. R. 2017. *#Republic: Divided Democracy in the Age of Social Media*. Princeton: Princeton University Press.
Sunstein, C. R. 2019. *Conformity: The Power of Social Influences*. New York: New York University Press.
Treier, S., and D. S. Hillygus. 2009. "The Nature of Political Ideology in the Contemporary Electorate." *Public Opinion Quarterly* 73 (4): 679–703.
Turner, J. C. 1991. *Social Influence*. Bristol: Open University Press.
Verba, S., K. Lehman Schlozman, and H. E. Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge: Harvard University Press.
Wardhaugh, R. 2015. *An Introduction to Sociolinguistics*, edited by J. M. Fuller. Blackwell Textbooks in Linguistics. Oxford: John Wiley & Sons.
Weeks, B. E., A. Halversen, and G. Neubaum. 2023. "Too Scared to Share? Fear of Social Sanctions for Political Expression on Social Media." *Journal of Computer-Mediated Communication* 29 (1): zmad041.

Cite this article: Schulz, W. Small (2025). What Would You Say? Estimating Causal Effects of Social Context on Political Expression. *Political Analysis*, 122–139. https://doi.org/10.1017/pan.2024.15