Appendix to “What Would You Say?” (2024)

William Small Schulz

A Sample Descriptives

The following pages provide sample descriptives for Studies 1-5.

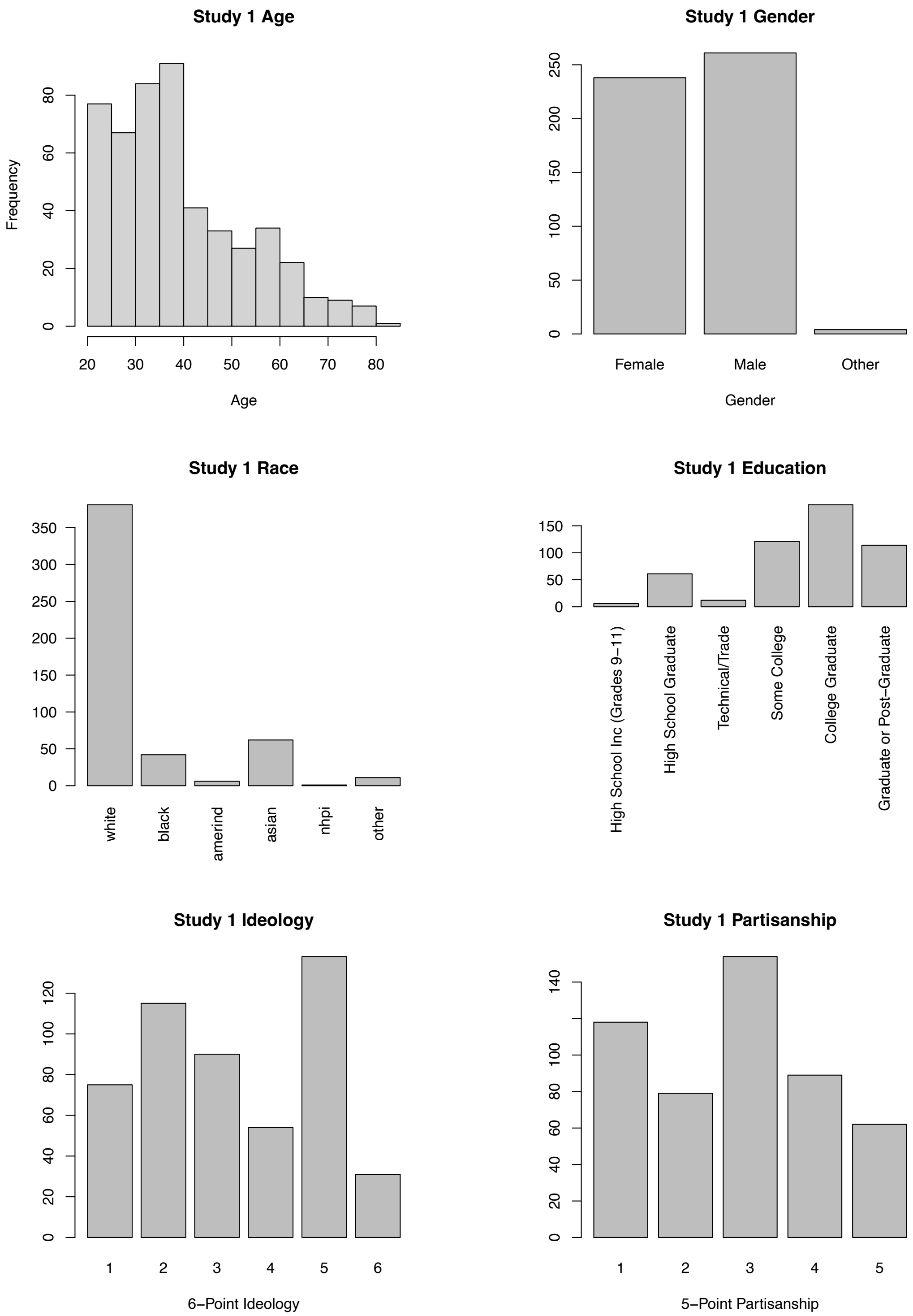


Figure 1. Study 1 Sample Descriptives

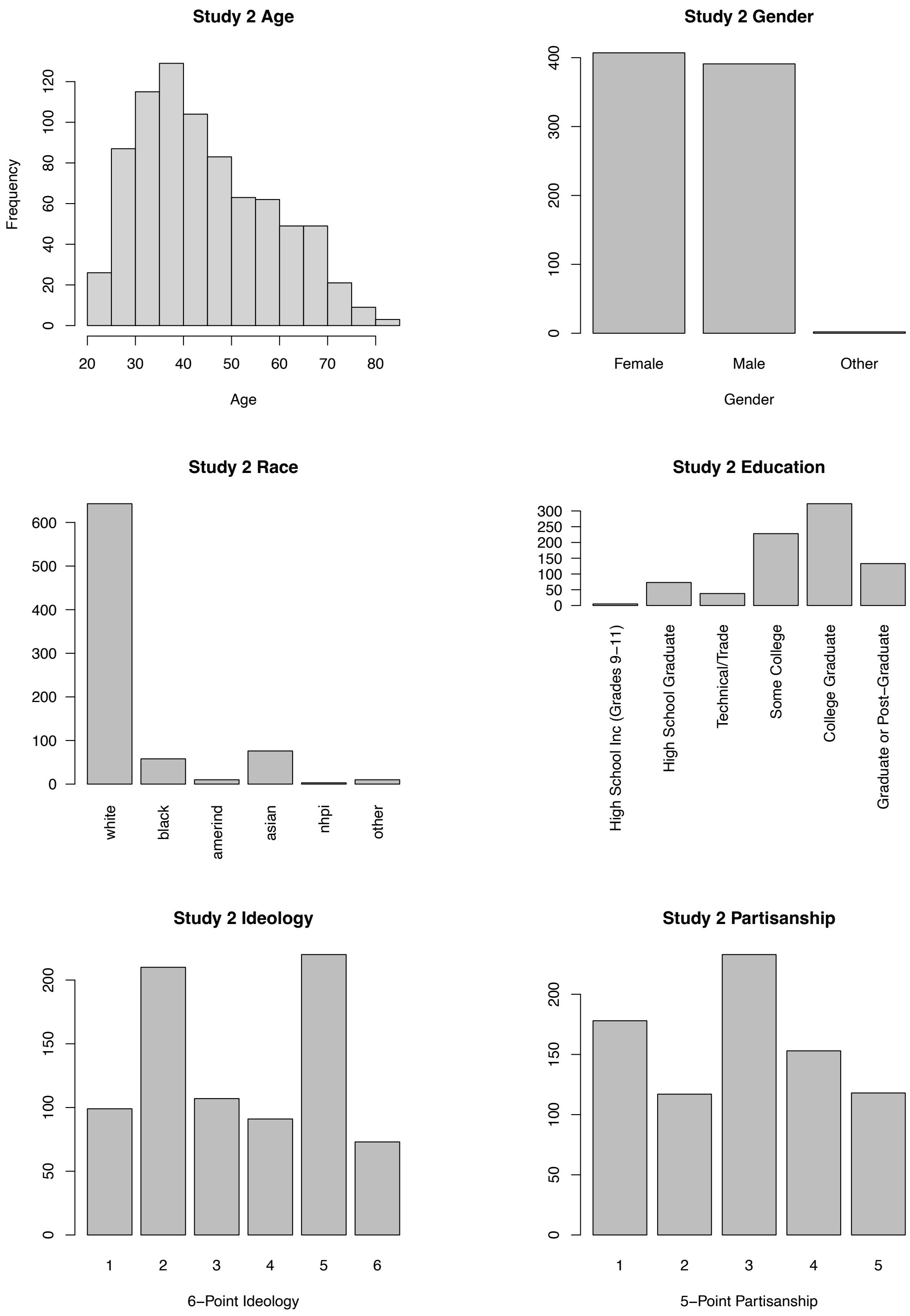


Figure 2. Study 2 Sample Descriptives

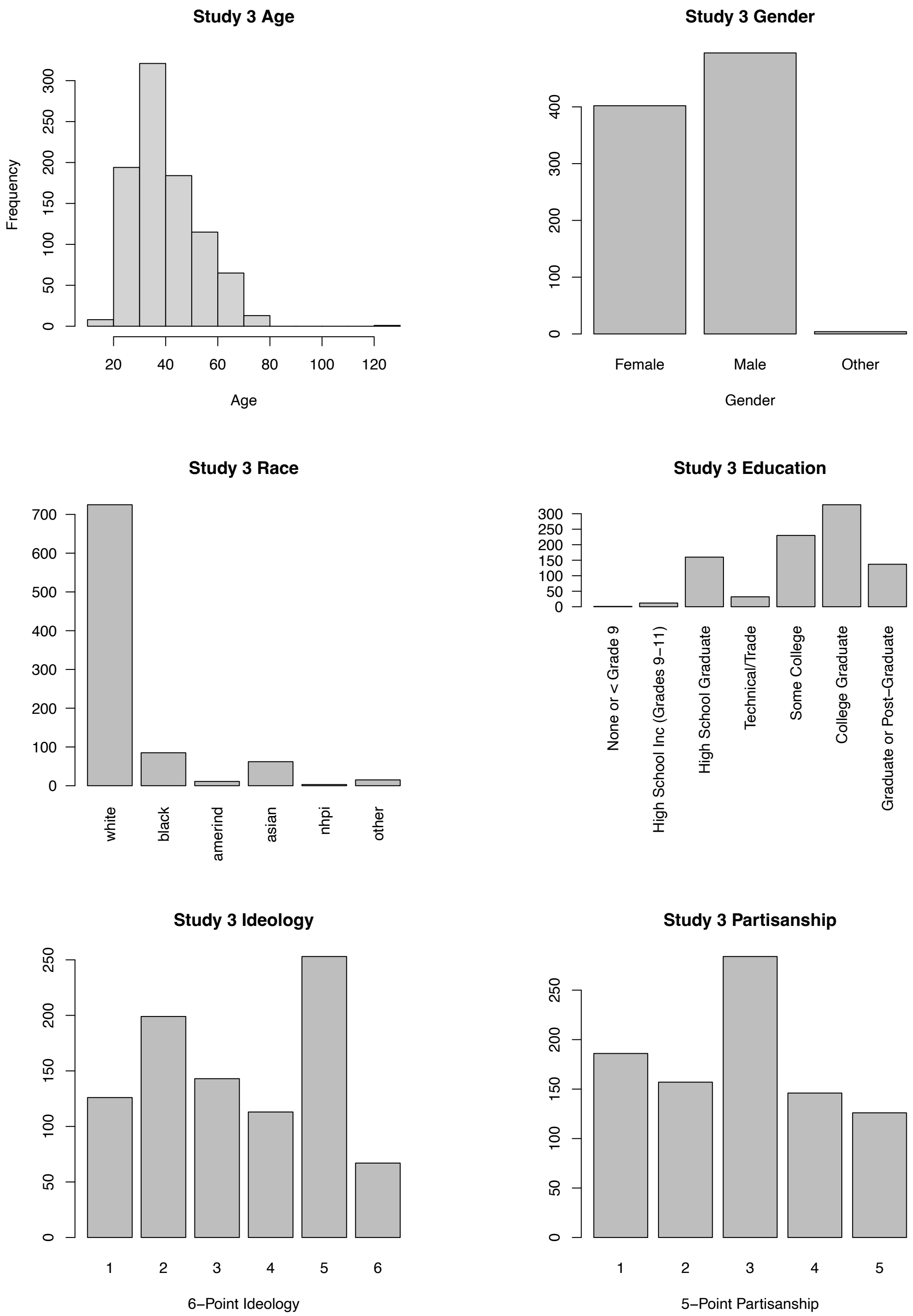


Figure 3. Study 3 Sample Descriptives

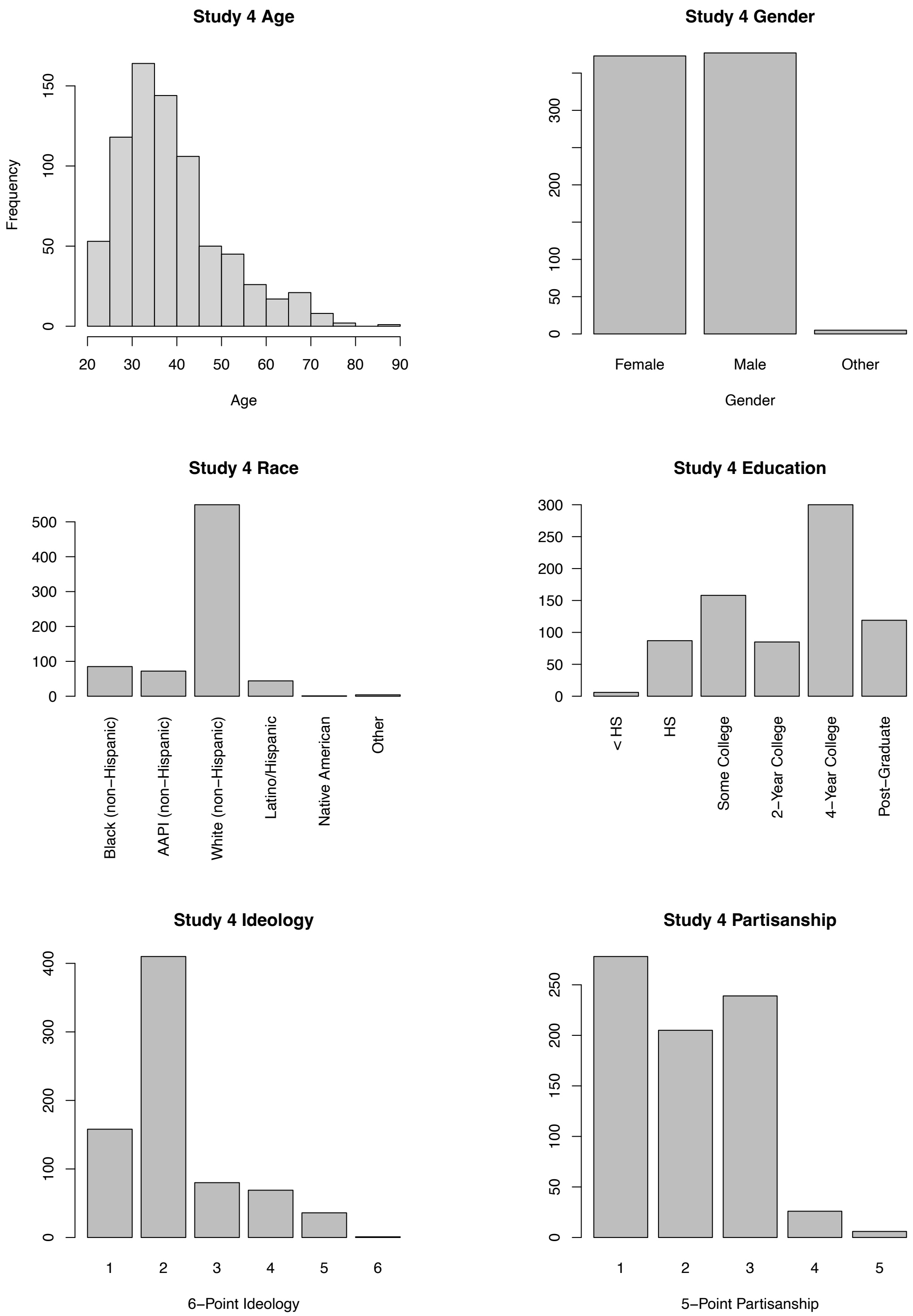


Figure 4. Study 4 Sample Descriptives

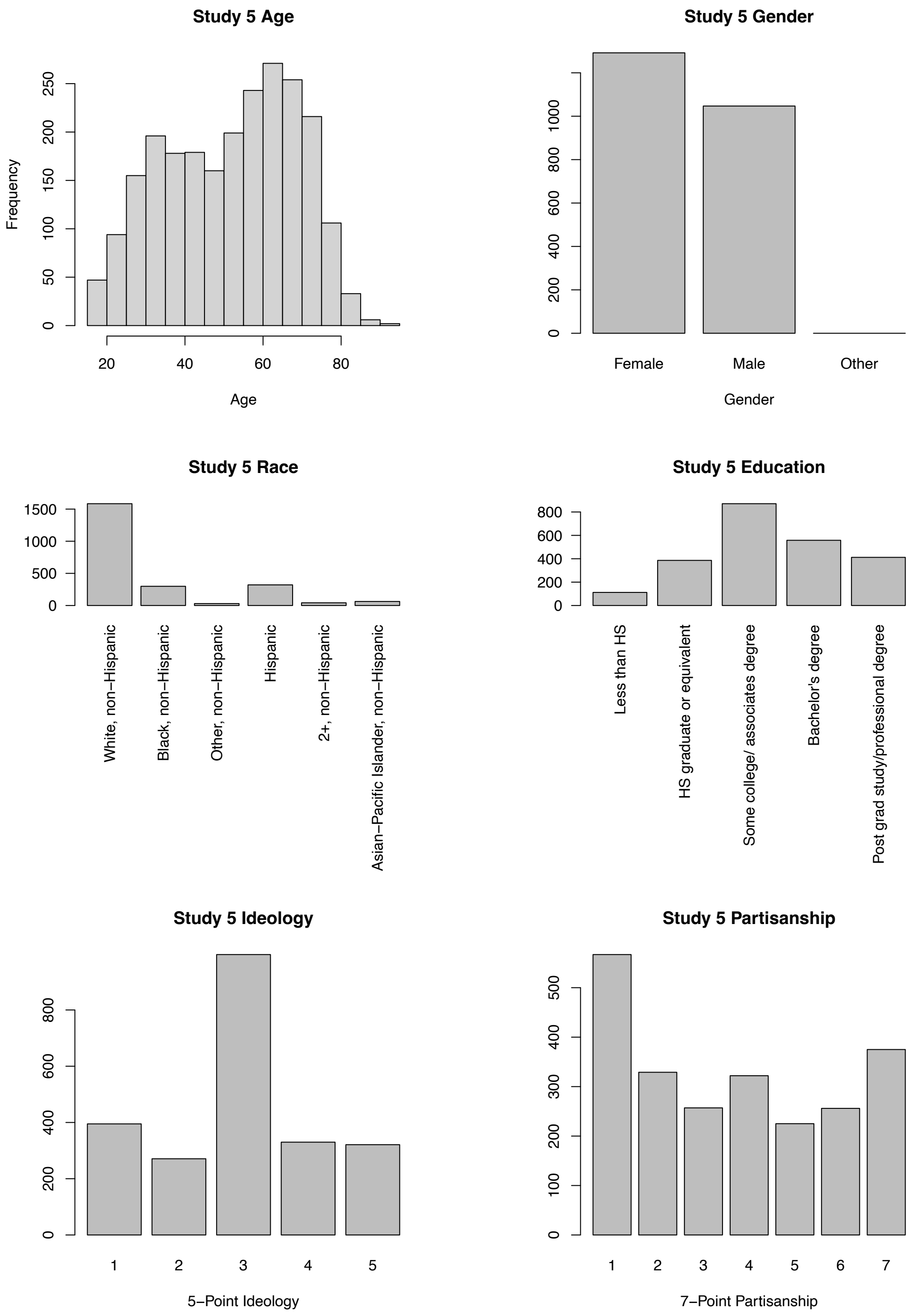


Figure 5. Study 5 Sample Descriptives

B Phrase Selection

To develop the set of words and phrases used in this study, I applied a mixture of qualitative and quantitative methods. Throughout the process, I sought to identify a set of phrases that varied in terms of topic, salience, technicality, ideological slant and extremity, and the relative literalness/implicitness of their political meaning. A mixed-methods approach furnishes a phrase selection procedure that was both theory- and data-driven, and also helps to illustrate the differences between my approach and a more conventional automated document scaling procedure.

Qualitatively, I compiled a list of phrases that I heard in the news media and on social media, and in casual conversations that I participated in or overheard. I included well-known political slogans from major contemporary political movements, such as MAGA and BLACK LIVES MATTER. I informally queried friends, colleagues, and audience members at presentations of my work-in-progress, to get suggestions of phrases to include. I specifically included ESTATE TAX and DEATH TAX since these terms were highlighted in (Gentzkow and Shapiro 2010) as being ideologically-diagnostic in congressional floor speeches. I also solicited phrase suggestions from participants in Studies 1 and 2, some of which I implemented in Study 3.

I also incorporated quantitative insights from a separate research project, in which I applied a lasso-regularized logistic regression model to a dataset of political tweets, and identified terms that were preferentially used by liberal and conservative Twitter users. To illustrate this, I present an analysis (that can be reproduced using the data and code posted in the online appendix) of a set of 1 million tweets, which were posted primarily in 2020 and 2021 (though it includes some earlier tweets) by a set of users who were identified as liberal and conservative (by a combination of methods including manual inspection, analysis of tweet texts, and Barberá's (2015) following-based *Tweetscores* method). The set of users included 260 liberals and 217 conservatives, but the tweet dataset included equal numbers of liberal and conservative tweets (500,000 each, for a total of 1 million). The data collection approach over-sampled especially prolific users, hence the high tweet-to-user ratio.

To analyze this dataset, I first pre-processed the tweets by removing punctuation and stopwords, constructed bigrams, and trimmed rare tokens and tokens used only by a small number of users. I also excluded terms beginning with , on the basis that mentions of Twitter users would not provide useful inspiration for phrases to include in the WWYS measure. With the resulting document-feature matrix (containing 1 million documents and 22,781 features), I used the R package `glmnet` to estimate a lasso-regularized logistic regression model to predict tweets' ideology labels from their text features, using 10-fold cross-validation to optimize the λ regularization parameter. The model with the optimal λ produced 18,588 non-zero feature weights, which was not amenable to direct inspection. To make the results tractable, I increased the λ penalty to bring more term weights to zero, to inspect only the most informative term weights.

However, there are limits to this approach, because the words and phrases that are most *predictive* of Twitter users' ideology are not necessarily words and phrases that are *semantically* ideological. A model with a relatively high penalty of $\lambda = .015$, for example, produced 79 non-zero term coefficients, which are plotted in Figure 6: although it is not surprising that these terms are predictive of ideology, the reasons that they are predictive are heterogeneous. True, several of the features selected by the model are ideologically-coded in the manner that my study seeks to explore (such as VOTER SUPPRESSION, which scales left, and PATRIOT, which scales right), but many of the terms have no symbolic interest. For example, the fact that TRUMP and REPUBLICAN scale to the left and BIDEN and DEMOCRAT scale to the right likely reflects that politically-interested Twitter users talk more about their opponents than about their allies (the fact that PRESIDENT TRUMP scales to the right has perhaps some symbolic interest, as this honorific bigram denotes respect, and therefore is more likely to be used by allies). Similarly, terms associated with religion and particularly Christianity are prevalent on the right, but the reasons for this are not very interesting

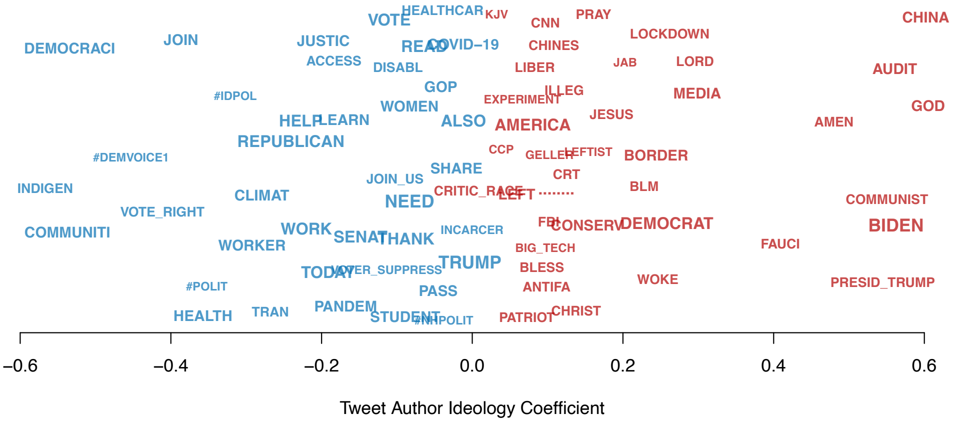


Figure 6. Coefficient plot at $\lambda = .015$. Based on tweets collected as part of Study 4, primarily authored in 2020-2021.

from a linguistic perspective. On the other hand, some terms plausibly reflect differences in political organization or political psychology between the left and the right (particularly in terms like NEED and HELP that scale to the left), which might be too sub-conscious to be desirable for inclusion in the WWYS question – although I ultimately chose to include some terms like this, to explore the possibility.

This exercise helps illustrate the distinction between the WWYS approach to studying lexical ideology, and a more conventional automated approach, by drawing our attention to the distinction between features that are *de facto* ideological in an automated analysis such as this (which can simply reflect that they come up more often in the topics that a certain group often discusses), and features that are ideological in the sense that individuals would prefer, all else equal, to use or avoid using certain terminology because of its ideological meaning. Because the WWYS question collects data by asking whether a person would use a word or phrase in the abstract, it is capable of isolating aspects of language that are socially-constructed to be ideological, such that individuals have ideological preferences about whether to use or avoid different terms.

To derive insights from this dataset without allowing terms with raw predictive power to crowd out terms with interesting ideological symbolism, I exported the term coefficients from a model with a more lenient λ penalty into a text file, and read the terms manually to identify candidates for inclusion. In an iterative procedure, I developed a list of features, and then searched the full set of feature weights to identify variations on related themes. Figure 7 plots the term coefficients of 38 features selected in this manner, many of which were implemented directly in the WWYS question.

Although I also included some phrases that were not extracted from the tweet dataset by this procedure, I used the tweet data as a source of inspiration, and as a guide to ensure that I included a variety of phrases that do in fact distinguish liberals from conservatives (at least those who use Twitter). I would recommend this approach to any future researchers who wish to update or adapt my method to a new setting, since it provides an empirical basis for identifying candidate phrases, while allowing the researcher to maintain control over the substantive focus of their study, and to apply their own knowledge and qualitative insights to distinguish "symbolism" from "mere predictive power" of phrases.

Future studies may wish to focus on a specific topical domain, but for my initial demonstration studies I sought to cover a variety of phrases to maximize my studies' exploratory potential. For Studies 1 and 2, I ultimately selected 46 words and phrases, evenly split between terms I expected to be left-slanted and terms I expected to be right-slanted, as shown in Table 1.

Many of the phrases can be seen as pairs, some of which are left- and right-slanted alternatives

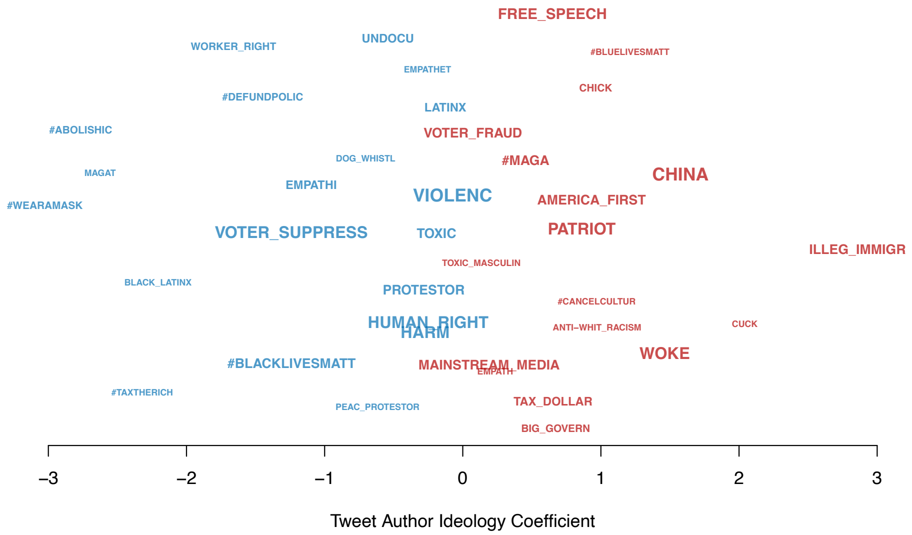


Figure 7. Coefficient plot of 38 manually-selected features from the lasso model. Based on tweets collected as part of Study 4, primarily authored in 2020-2021.

for referring to the same thing (such as UNDOCUMENTED IMMIGRANT and ILLEGAL IMMIGRANT), others that refer to alternative framings or topical foci within a broader issue domain (such as VOTER SUPPRESSION and VOTER FRAUD), or alternative slogans (like BLACK LIVES MATTER and ALL LIVES MATTER/BLUE LIVES MATTER), as well as others that are not direct alternatives to each other but share a general political theme (like MANSPLAIN and CHICK, which both pertain to gender, and DOG WHISTLE and FREE SPEECH, which both pertain to speech). I did this primarily because I considered these pairings to be a substantively interesting aspect of lexical ideology – I did also include some terms that are not intended to be paired (such as PRIVILEGE, CANCEL CULTURE, WEAR A MASK, and BIG GOVERNMENT). Among the terms that are not topically-paired, there are some similarities in *types* of terms (for example, WHITE TRASH and CUCK are both pejorative, and HARM and STRONG, which were identified in the tweet analysis, both connote general values that might have conscious ideological symbolism, or might reflect sub-conscious differences in political psychology.

The inclusion of paired terms also emerged somewhat naturally: many politically-charged phrases are constructed to be alternatives to existing terminologies (like LATINX as an alternative to LATINO, or DEATH TAX as an alternative to ESTATE TAX), and even when these constructions are not deliberate it is often the case that language acquires political symbolism when ideologues have different ways of talking about the same topic. It is likely that a future researcher seeking to reproduce my procedure would be able to find many paired terms like these, and I would recommend this as a way to achieve a general topical balance in the stimuli, which I deem substantively desirable even if it is not strictly necessary.

For Study 3, I used an updated set of 40 phrases (see Table 2). The updated set had considerable overlap with the original set, but dropped some of the original terms in order to include some new phrases. New phrases were based on colleague's requests, personal interest, and suggestions solicited from participants in Studies 1 and 2. I did not seek to include paired terms in this study, but simply sought to explore additional phrases while retaining the general themes of the phrases in Studies 1 and 2. Study 4 employed a briefer version of the WWYS question, based on the phrases used in Studies 1 and 2 (see Appendix K for details).

Table 1. Phrases Used in Studies 1 and 2

Expected Left-Slanted	Expected Right-Slanted
SYSTEMIC RACISM	REVERSE RACISM
LATINX	LATINO
ESTATE TAX	DEATH TAX
UNDOCUMENTED IMMIGRANT	ILLEGAL IMMIGRANT
WORKERS' RIGHTS	RIGHT TO WORK
VOTER SUPPRESSION	VOTER FRAUD
MANSPLAIN	CHICK
DOG WHISTLE	FREE SPEECH
POST-TRUTH	MAINSTREAM MEDIA
MICRO-AGGRESSION	CHINA VIRUS
PROTESTER	RIOTER
SILENCE IS VIOLENCE	THUG
BLACK LIVES MATTER	ALL LIVES MATTER
ABOLISH THE POLICE	BLUE LIVES MATTER
DEFUND THE POLICE	AMERICA FIRST
ACAB	MAGA
PRIVILEGE	CANCEL CULTURE
WEAR A MASK	BIG GOVERNMENT
HUMAN RIGHTS	PATRIOT
WHITE TRASH	CUCK
EMPATHY	SNOWFLAKE
TOXIC	WOKE
HARM	STRONG

Table 2. Phrases Used in Study 3

Expected Left-Slanted	Expected Right-Slanted
SYSTEMIC RACISM	BIG GOVERNMENT
DEFUND THE POLICE	AMERICA FIRST
HETERONORMATIVE	ALL LIVES MATTER
MANSPLAIN	REVERSE RACISM
MICROAGGRESSION	VIRTUE SIGNAL
SAFE SPACE	ILLEGAL ALIEN
EAT THE RICH	TRADITIONAL VALUES
PRIVILEGE	CRITICAL RACE THEORY
WORDS MATTER	LIBTARD
CLIMATE CRISIS	PERSONAL RESPONSIBILITY
LATINX	CANCEL CULTURE
POC	MAINSTREAM MEDIA
HUMAN RIGHTS	SNOWFLAKE
TOXIC MASCULINITY	PATRIOT
EQUITY	BITCH
GLOBAL SOUTH	THIRD WORLD
CIS-GENDER	PROSTITUTE
TRIGGERED	BIOLOGICAL WOMEN
ABORTION IS HEALTHCARE	SANCTITY OF LIFE
EMPATHY	DO YOUR OWN RESEARCH

C Wordsticks Model

In order to estimate lexical ideology and outspokenness 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 lexical ideology, modeled as a normally-distributed latent trait α_i.
- Probability of saying each phrase j is either strictly increasing or strictly decreasing in lexical ideology, such that for each phrase j, two cutpoints $c_j^{[1]}$ and $c_j^{[2]}$ can exhaustively partition the lexical 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. Corresponding assumptions apply to the 4-point response scale implemented in Study 3.
- Response probabilities can be modeled as an inverse logit transformation of a continuous utility function. This can be interpreted as an assumption that responses are generated by deterministic utility maximization after a random utility shock that is distributed as type 1 extreme value, however it is more forthright to say that an inverse logit function is a convenient and reasonable way to map from the continuous latent utility to the ordinal observed outcome.
- The direction and rate of change in probability of saying each phrase j, as a function of respondent lexical ideology α_i, varies across phrases, and can be summarized with a scalar γ_j, which multiplies the distance between the respondent's lexical 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 lexical 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 lexical 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,

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

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 ideal point
$\gamma_j \in \mathbb{R}$ 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 cutpoints

Note also the following rearrangement:

$$\begin{aligned}
\mu_{ij}^{[k]} &= (\alpha_i - c_j^{[k]}) \times \gamma_j + \beta_i && \text{(A.2)}\\
&= \alpha_i \times \gamma_j - c_j^{[k]} \times \gamma_j + \beta_i && \text{(A.3)}\\
&= \alpha_i \times \gamma_j + \beta_i - c_j^{[k]} \times \gamma_j && \text{(A.4)}\\
&= \alpha_i \times \gamma_j + \beta_i - c_j^{[k]\prime}, \quad \text{where } c_j^{[k]\prime} = c_j^{[k]} \times \gamma_j && \text{(A.5)}
\end{aligned}$$

This formulation makes clear that Wordsticks is a straightforward ordinal extension of Wordfish, as originally specified by Slapin and Proksch (2008):

$$\Pr(y_{ij} = k) \;\; = \frac{(\lambda_{ij})^{[k]} \times \exp(-\lambda_{ij})}{k!}; \qquad \lambda = \exp(\mu'_{ij}); \qquad \mu'_{ij} = \alpha_i \times \gamma_j + \beta_i + \theta_j \tag{A.6}$$

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

The formulation in Equation A.5 is also more convenient to estimate in the STAN modeling language, as shown in Appendix D, where I define the phrase utility thresholds as:

$$c_j^{[k]\prime} = \sum_1^k d_j^{[k]}, \text{ where } d_j^{[k]} > 0 \text{ for all } k > 1 \tag{A.7}$$

which allows flexible estimation of the cutpoint locations in utility space, and renders the inclusion of a phrase-level intercept (as in Wordfish) redundant. This approach extends to any[5] ordinal response scale, and aids model identification by constraining all $c_j^{[k]} < c_j^{[k+1]}$, which avoids reflection invariance (Bafumi *et al.* 2005). This is necessary because in substantive terms, $c_j^{[1]}$ represents the point in utility space where "might" becomes more a probable response than "wouldn't" for phrase j, and $c_j^{[2]}$ is the point where "would" becomes more probable than "might" – exchanging $c_j^{[1]}$ for $c_j^{[2]}$ would be equivalent to flipping the sign of the phrase slant γ_j and the model would be unidentified. Additionally, I impose a standard normal prior on α and place a sign constraint on one of the phrase slants γ_j to identify the direction of the ideological dimension, and I place a normal prior on β with mean zero and flexible spread on the respondent outspokenness β. See Appendix D for model code.

[5] Studies 1, 2, and 4 used a 3-point scale and thus 2 cutpoints, while Study 3 used a 4-point response scale and thus required 3 cutpoints. See Appendix D for details.

D Estimation of Wordsticks Model

I implement Wordsticks in the STAN modeling language (Stan Development Team 2022), as implemented in the R package `RStan` (Stan Development Team 2023).

D.1 3-Point Model

The following model represents an implementation of Wordsticks where data are collected on a 3-point scale (wouldn't/might/would) as in Studies 1, 2, and 4.

D.2 4-Point Model

The logic of the above 3-point model can easily be extended to a 4-point response scale (definitely wouldn’t/probably wouldn’t/probably would/definitely would) as in Study 3.

D.3 Pooled 3- and 4-Point Model

In order to make the most of my available data, I estimated a pooled model including responses collected on a 3-point scale and those collected on a 4-point scale. The code for this model is printed below.

Figure 8 verifies that the estimates of α, β, and γ derived from the pooled model are nearly identical to those derived from models estimated separately on the 3-point and 4-point datasets. I therefore use estimates from the pooled model in all analyses in this paper, for parsimony.

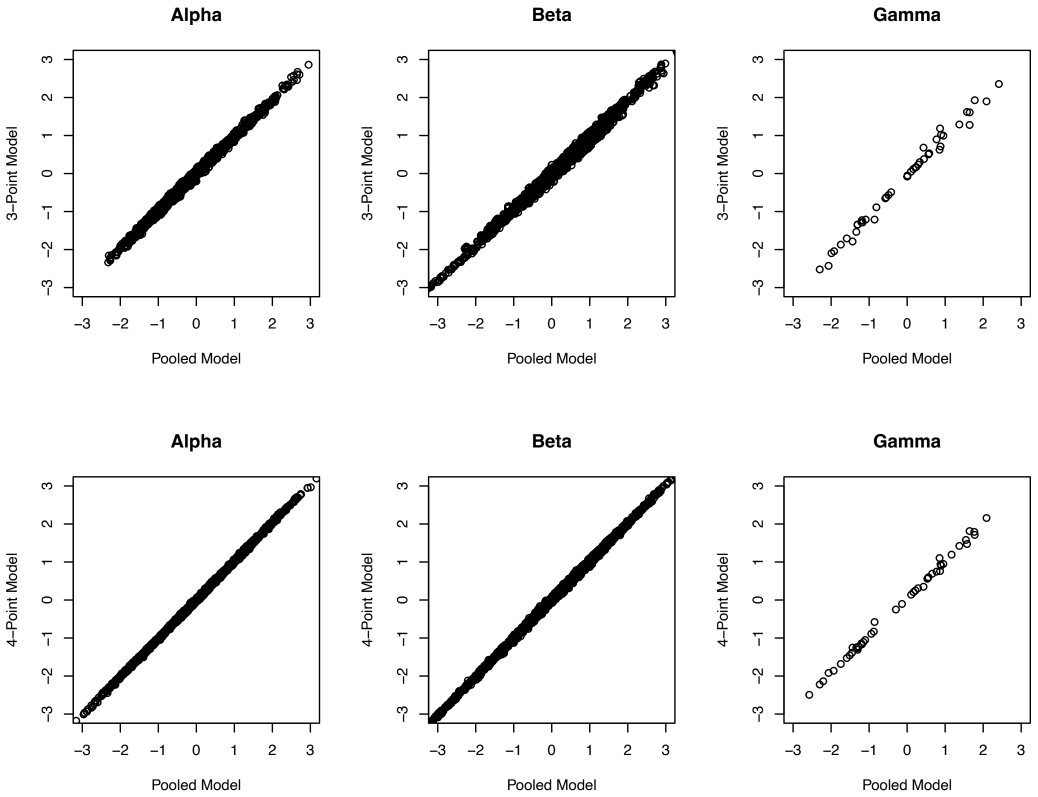


Figure 8. Scatterplots of alpha, beta, and gamma estimates derived from the pooled model (x axes) compared to the 3-point model (y axes, top row), and the 4-point model (y axes, bottom row). Based on data from all studies, collected 2021-2023.

D.4 Identification and Priors

I here describe identification constraints and priors used in the models whose code is shown above. To identify the direction and scale of the ideological dimension, I impose a standard normal prior on `alpha_raw`, which I hard-standardize as `alpha` in the transformed parameters block as recommended by Arnold (2018), and I constrain the phrase AMERICA FIRST to have a positive-signed slant by drawing its `gamma` from an exponential distribution with rate parameter equal to .1, in addition to additively constructing the cutpoints `c` in such a way that the upper cutpoint must always be greater than the lower cutpoint. Furthermore, I impose a normal prior on `beta` with mean zero and flexible standard deviation, to avoid additive invariance between `beta` and `c`.

In estimation I initialize `alpha_raw` at the respondents' standardized self-reported ideology on a 6-point likert scale, `beta` at the standardized row-means of the WWYS data matrix, and `gamma` at the standardized Pearson coefficients of correlation between WWYS responses and respondents' self-reported ideology on a 6-point likert scale. This initialization is not strictly necessary, but speeds model convergence. Diagnostic plots to visualize convergence and mixing are presented in Appendix E.

D.5 Train-Test Split

Egami *et al.* (2022) recommend a sample-splitting procedure for studies that conduct causal inference with latent variables, in order to avert violations of the single unit treatment value assumption (SUTVA) that arise when an outcome is measured using a model trained on the same data used to estimate causal effects. The authors therefore recommend training the measurement model on one partition of the available data, then applying the trained model to a held-out test set, and using only the latter dataset to estimate causal effects.

I therefore implemented a modified version of the Wordsticks estimation procedure described above. First, I estimated the Pooled 3- and 4-Point Model, exactly as presented in Appendix D.3. Then, I extracted the phrase parameters (`gamma`, `d3_A`, `d3_B`, `d4_A`, `d4_B`, and `d4_C`) estimated from this model, and estimated a second-stage model on the test set in which these parameters were held fixed at their estimated values from the first stage. The STAN code for the second-stage model (in which the first-stage phrase parameter estimates are read in in the data block) can be found on the following page.

In order to implement this procedure without undermining statistical power in the experiments I analyze in Studies 1 and 3, I used data from Studies 2, 4, and 5 as my training set, which allowed me to retain all observations from the experiments in Studies 1 and 3 in my test set. This approach breaks the dependence between the discovery of the phrase parameters and the estimation of causal effects in Studies 1 and 3, averting a SUTVA violation, without wasting any of the experimental observations.

This procedure also addresses concerns about overfitting of the model to detect a treatment effect. The fact that Wordsticks is directly inherited from Wordfish (Slapin and Proksch 2008) should mitigate concerns about *fishing* on the part of the analyst. Also, because feature selection is conducted *prior* to data collection, the WWYS method has fewer researcher degrees-of-freedom than typical text analyses. That said, it is possible that training the model on the experiment data could over-fit the model to dimensions of variation induced by the treatments. By withholding all data from the experiments in Studies 1 and 3 in the unseen test set, the sample-splitting procedure I implemented also prevents this specific form of over-fitting.

E Mixing Plots for Wordsticks Model

In this appendix I display mixing plots corresponding to the posterior samples for the latent variables `alpha`, `beta`, `gamma`, `d_1`, and `d_2`, estimated in the Stan implementation of the Wordsticks model (as described in Appendix D). In these plots, each color corresponds to one of three chains, and provides a useful visual diagnostic to ascertain whether the chains are well-mixed and fully explore the posterior of each parameter.

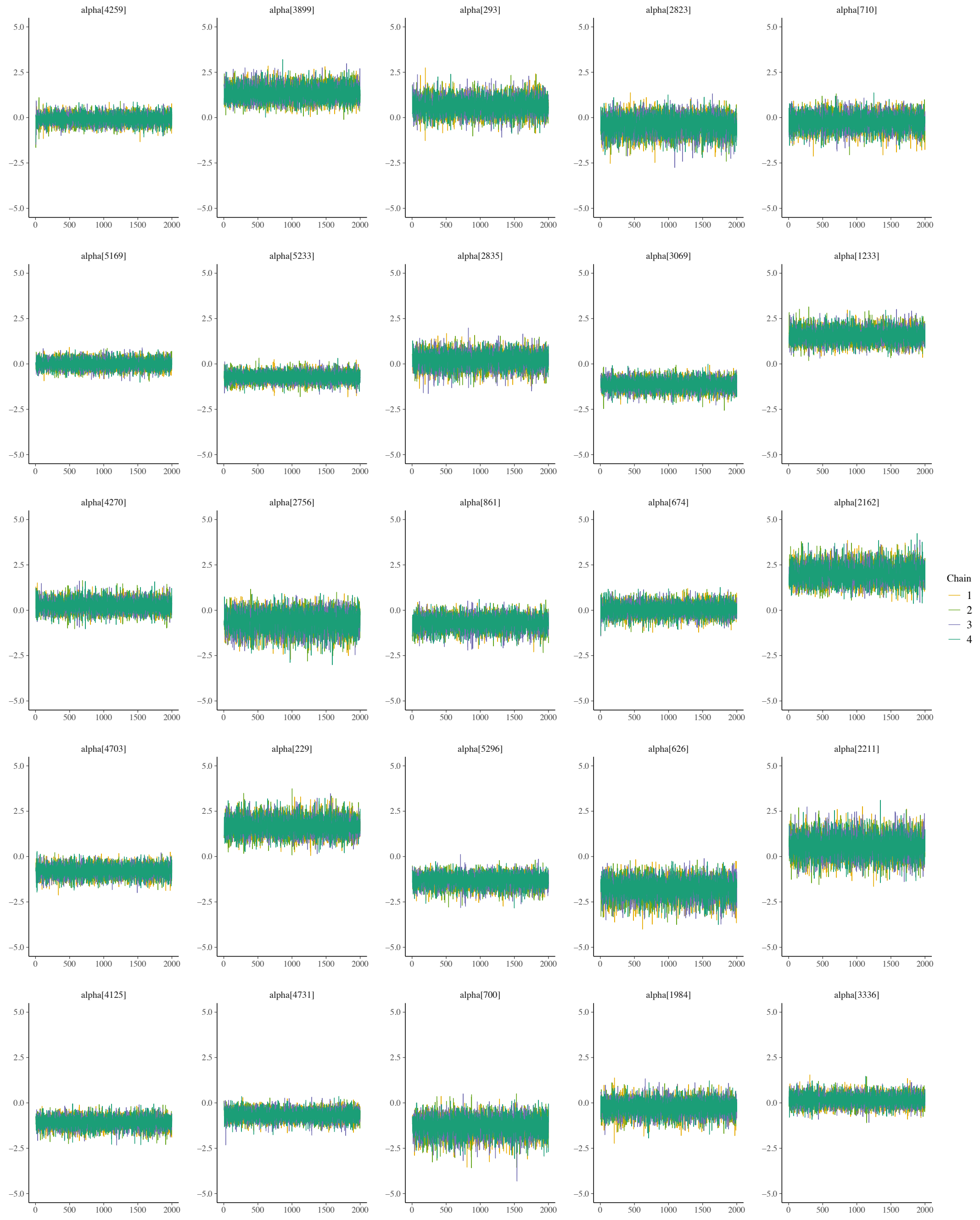


Figure 9. Mixing plot for respondent lexical ideal points α. A random sample of respondents are displayed (see online supplementary materials).

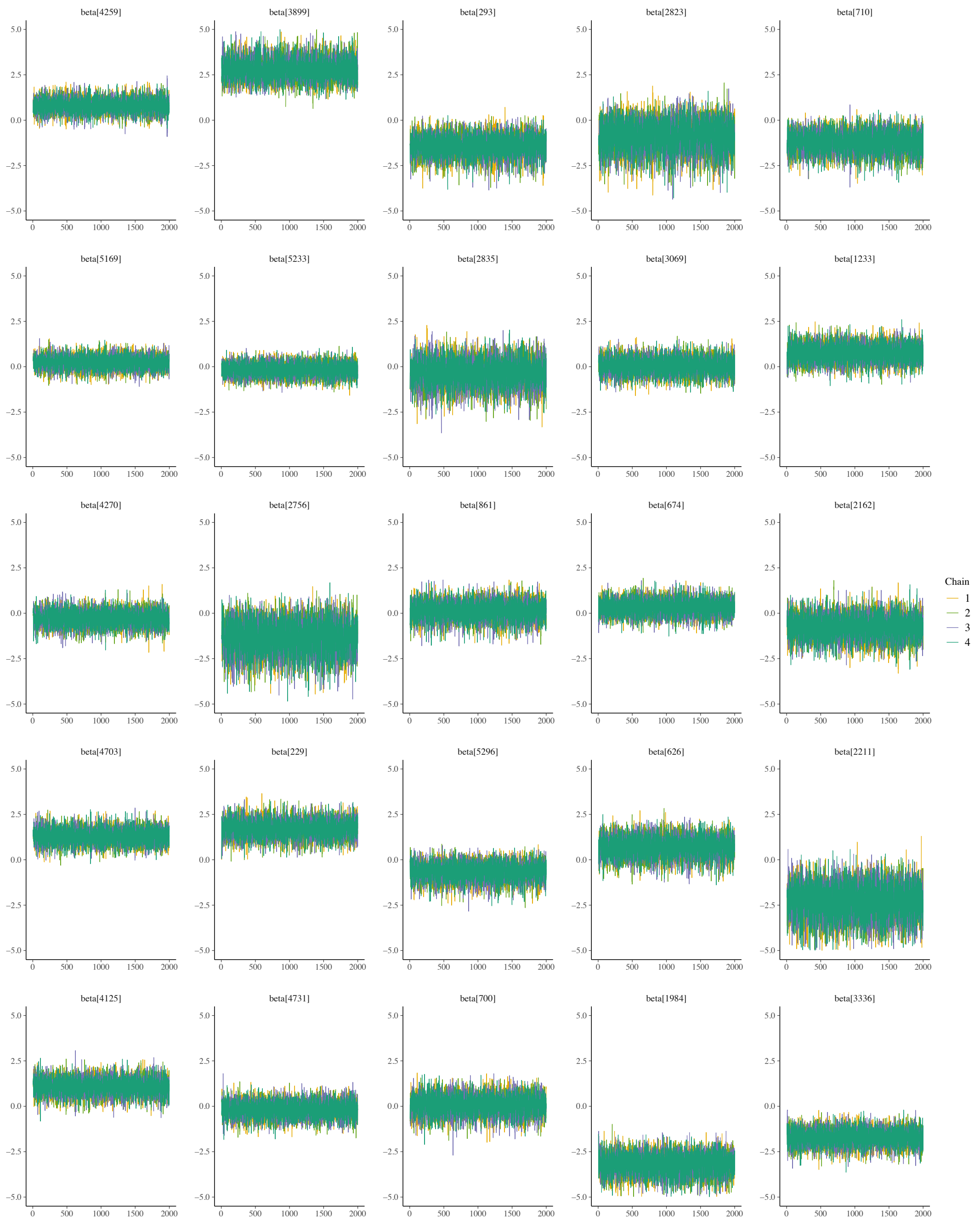


Figure 10. Mixing plot for respondent outspokennesses β. A random sample of respondents are displayed (see online supplementary materials).

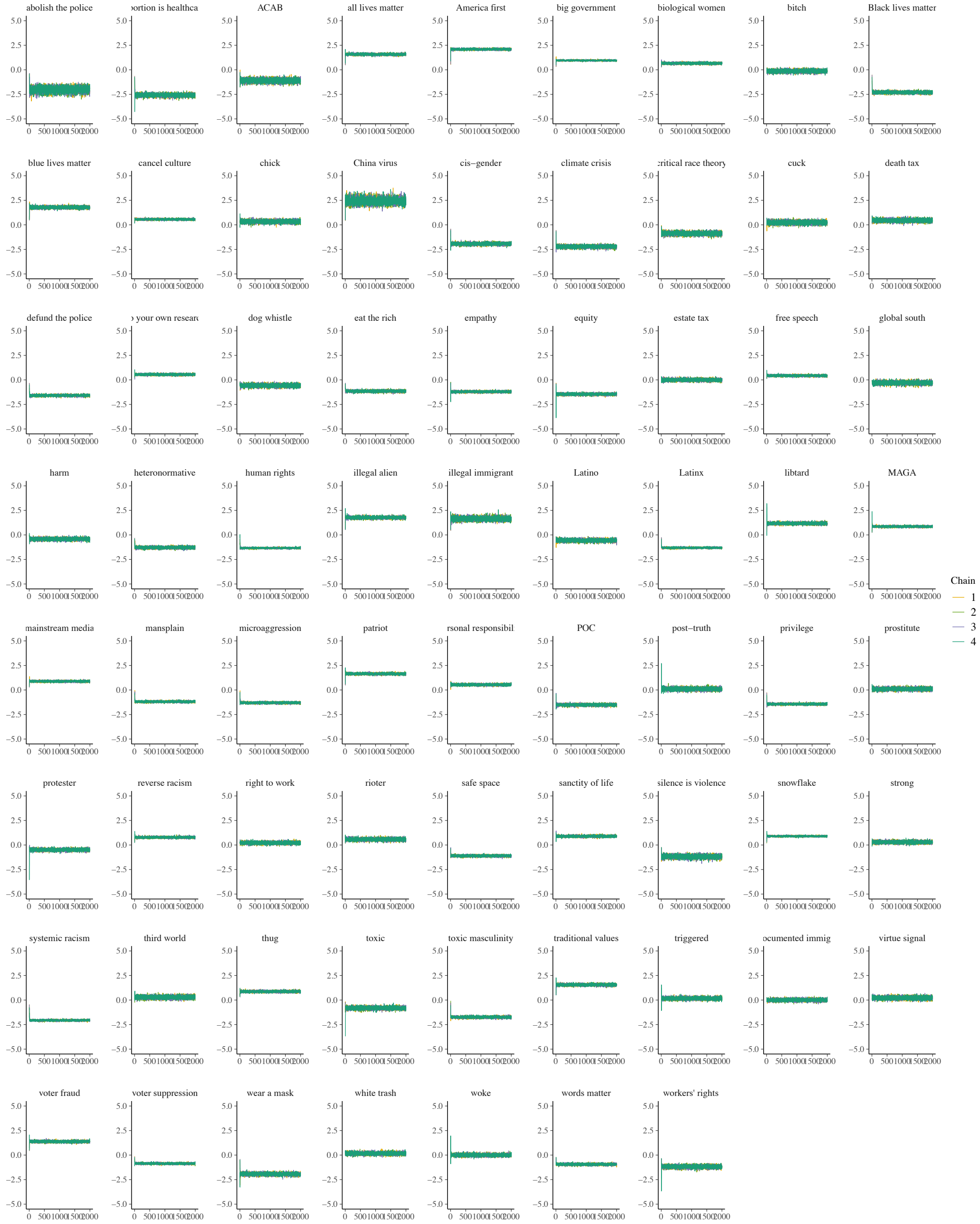


Figure 11. Mixing plot for item slants γ.

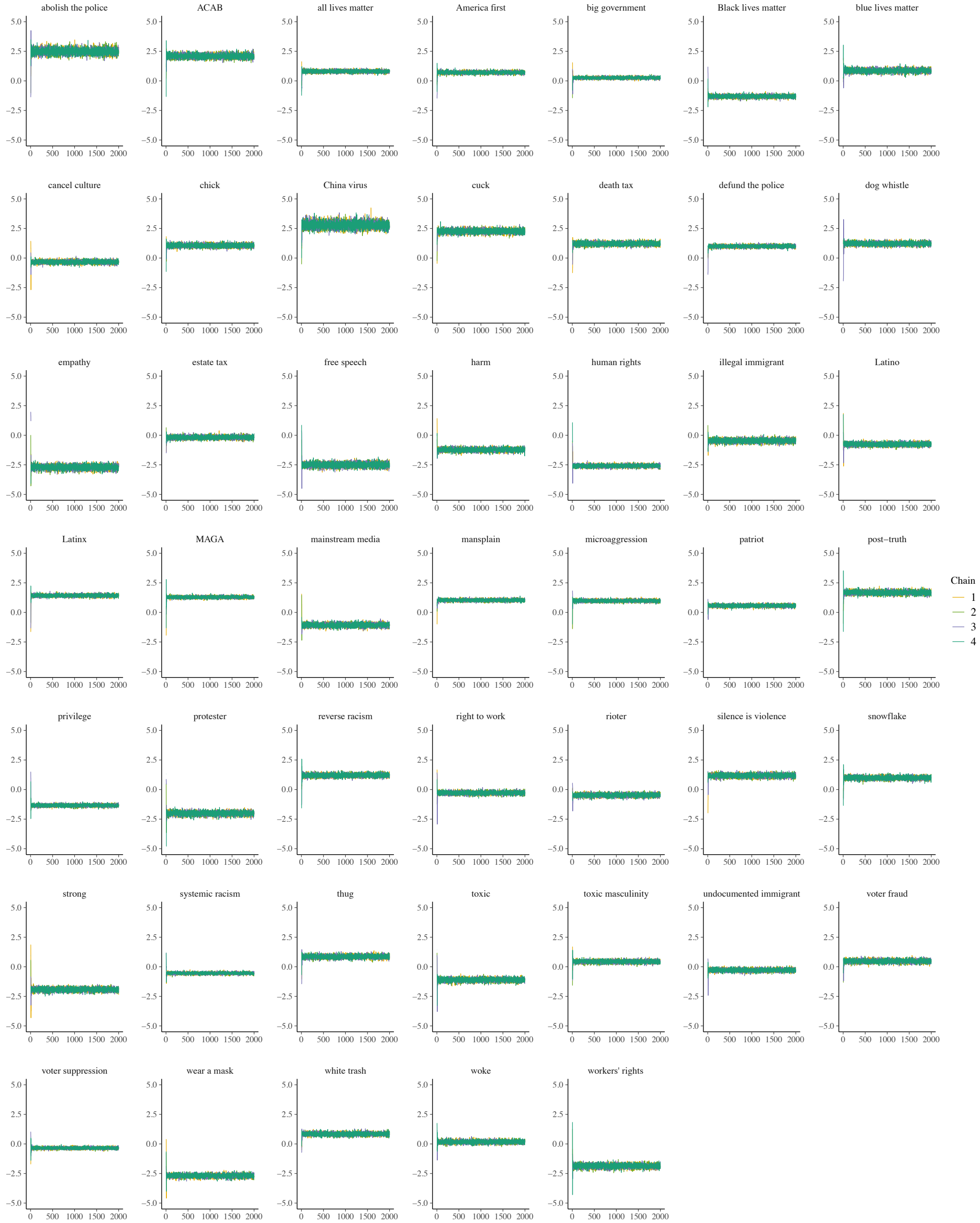


Figure 12. Mixing plot for item additive cutpoint component 1, for items with a 3-point response scale.

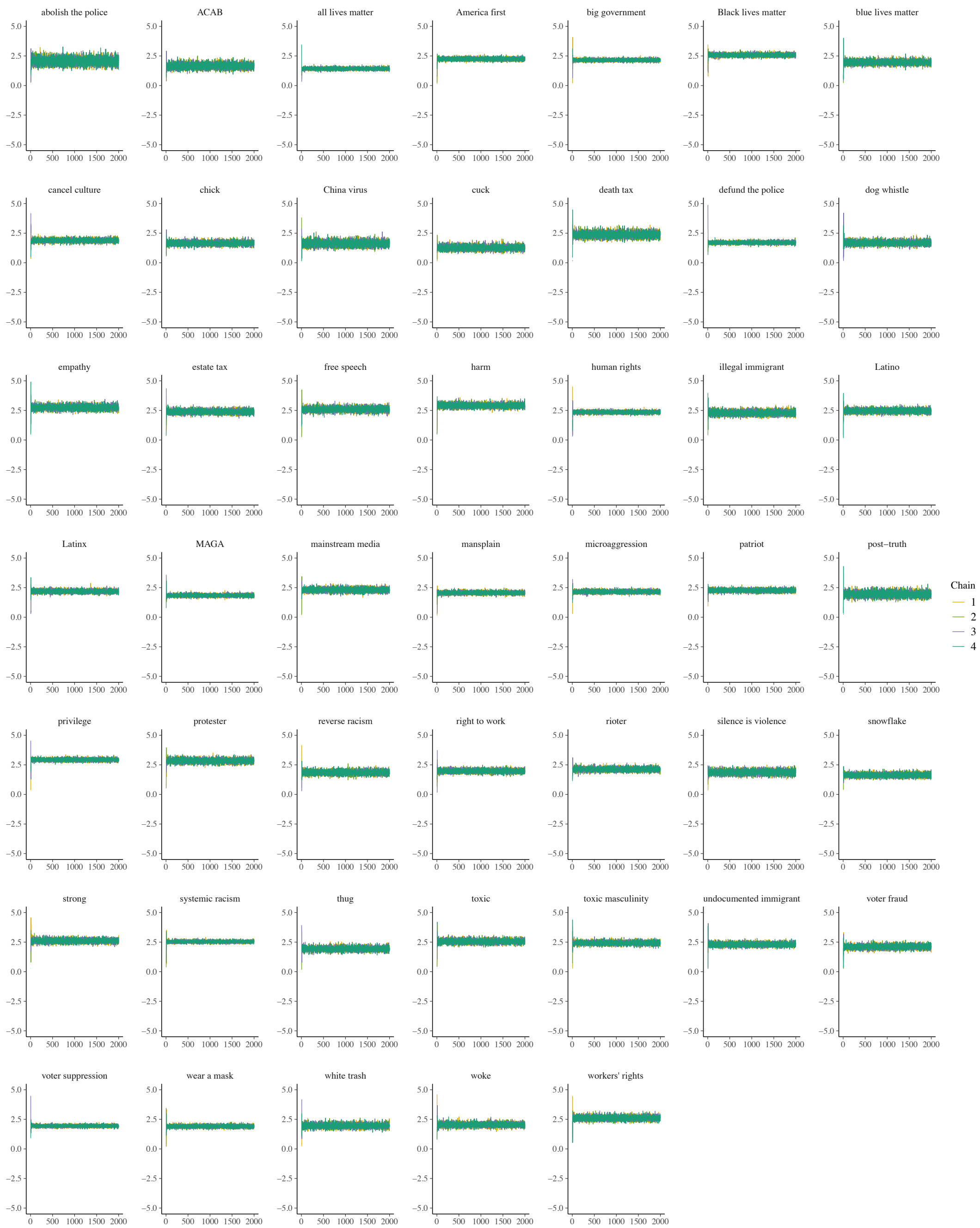


Figure 13. Mixing plot for item additive cutpoint component 2, for items with a 3-point response scale.

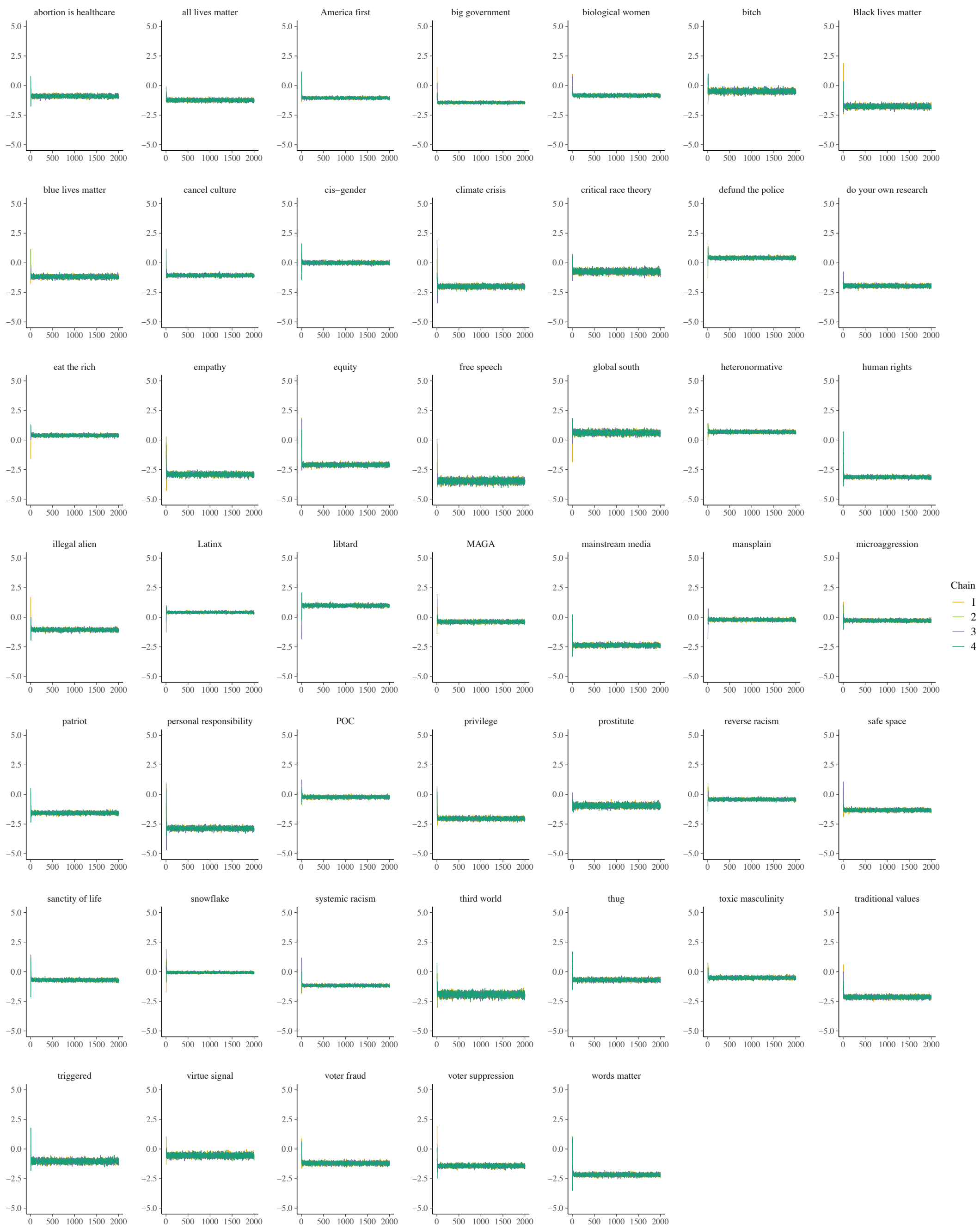


Figure 14. Mixing plot for item additive cutpoint component 1, for items with a 4-point response scale.

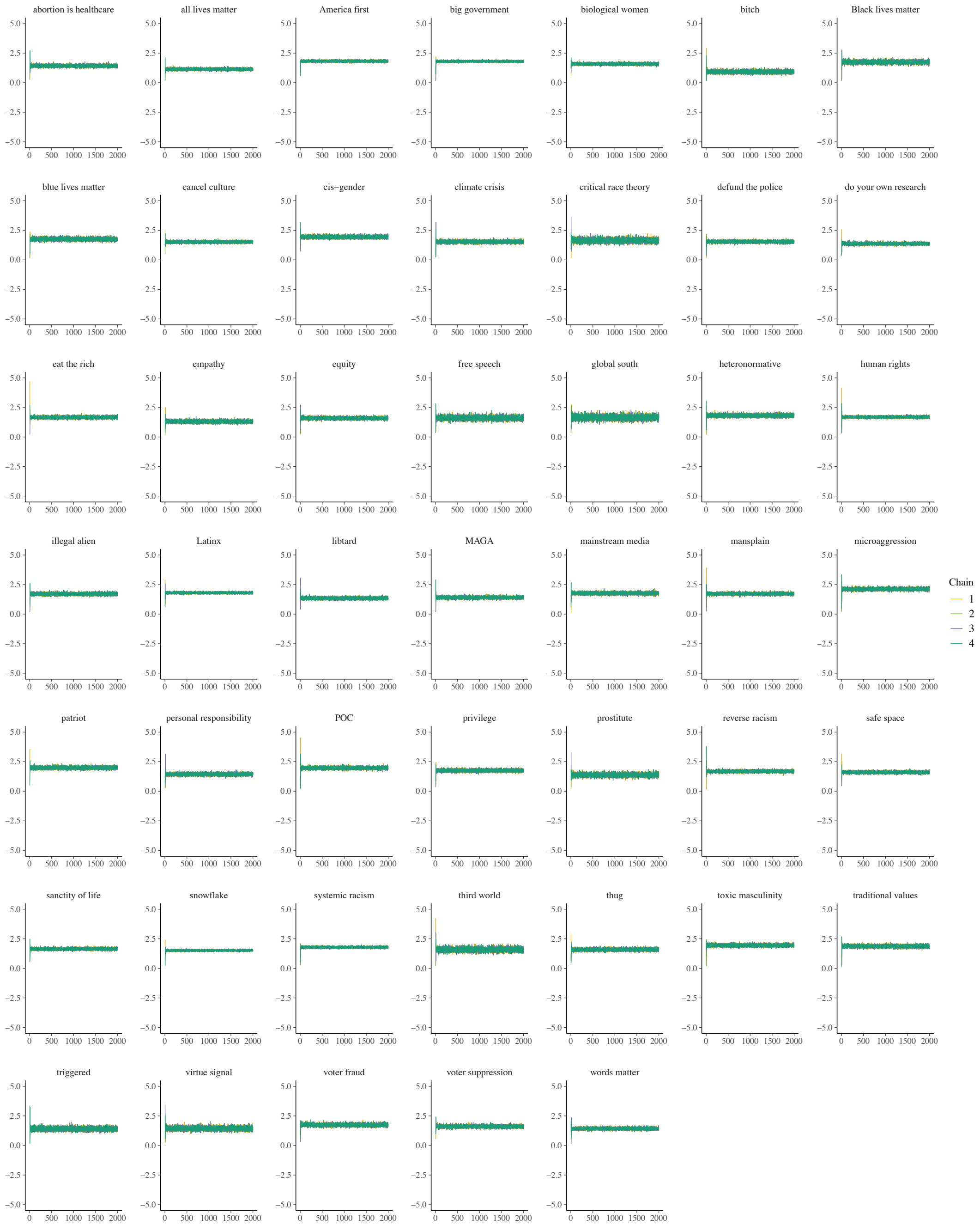


Figure 15. Mixing plot for item additive cutpoint component 2, for items with a 4-point response scale.

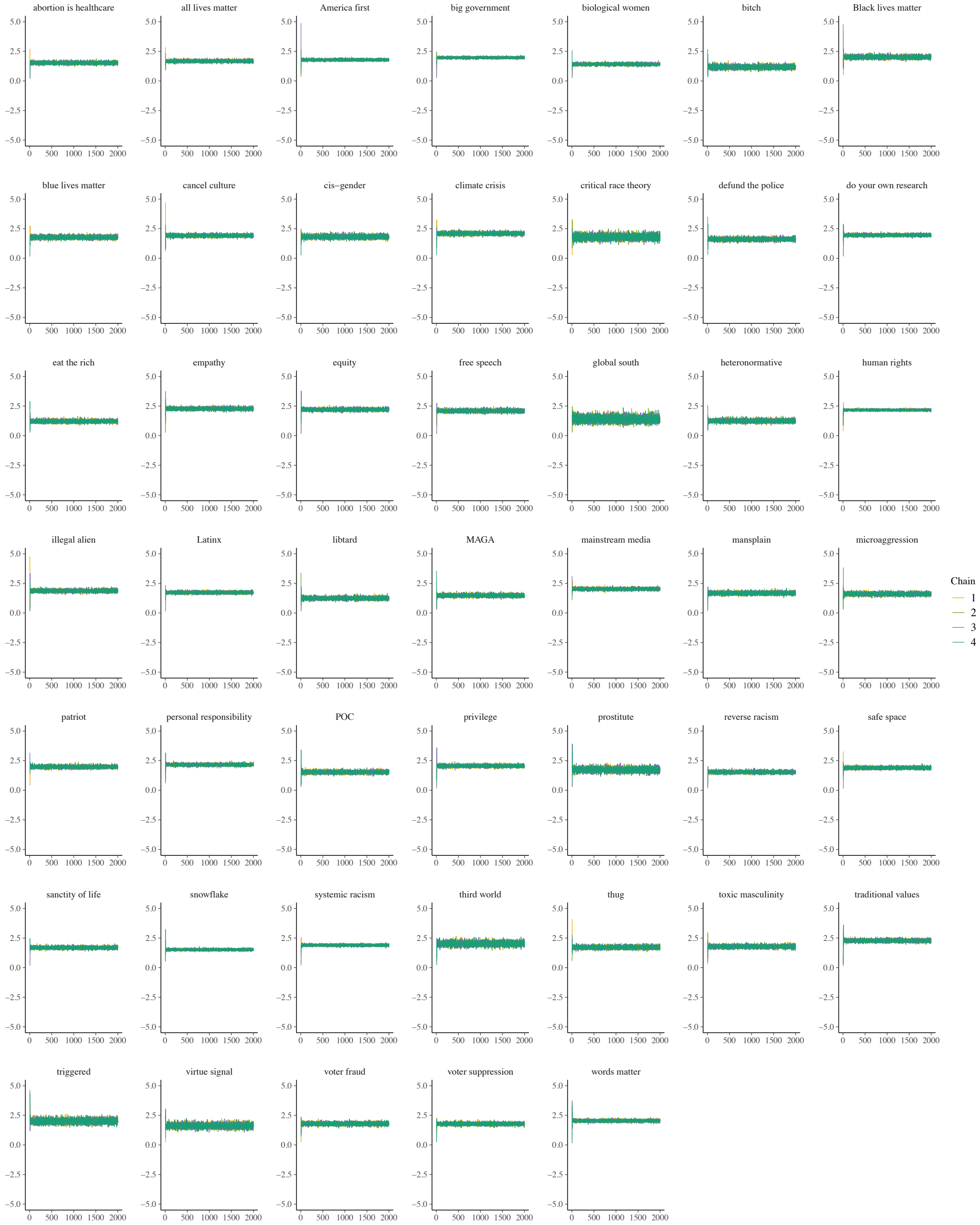


Figure 16. Mixing plot for item additive cutpoint component 3, for items with a 4-point response scale.

F Phrase Slants (Gammas)

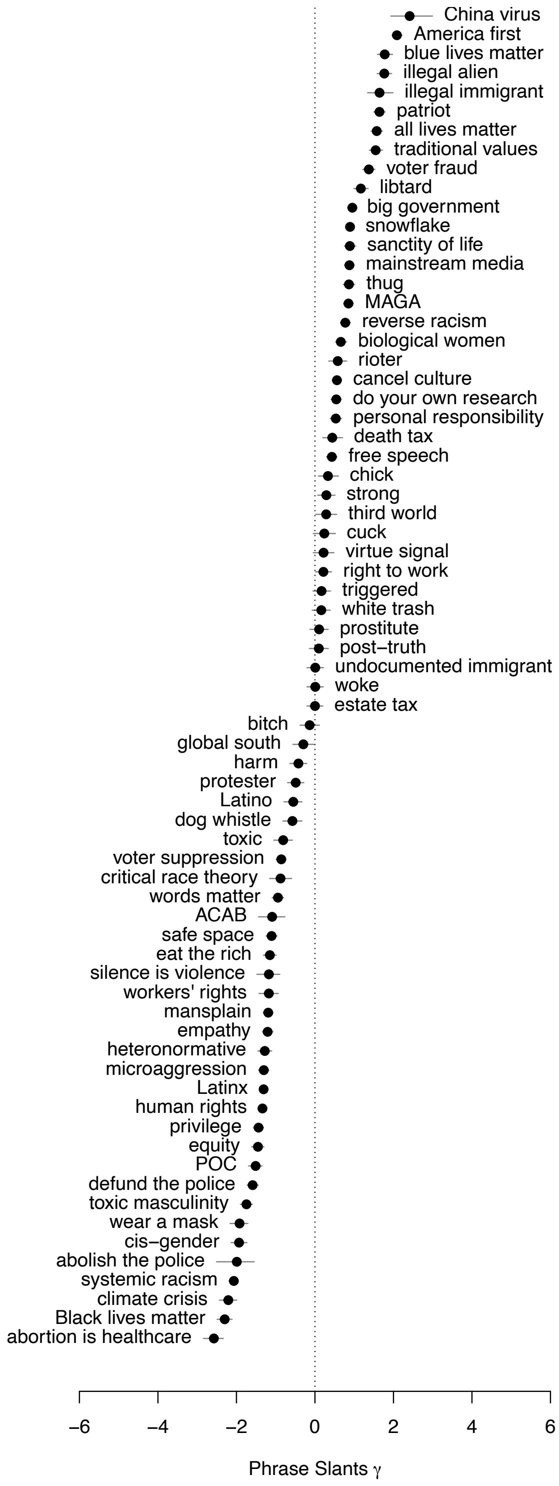


Figure 17. Gamma slant parameter estimates (points represent posterior medians, whiskers show 95% CIs) for all phrases, based on pooled model. Gammas represent phrases' slant or discrimination in the lexical ideology space. Phrases with large negative-signed gamma values are highly discriminative in 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 lexical ideology in this model. Based on data from all studies, collected 2021-2023.

G Bivariate Response Scatterplots

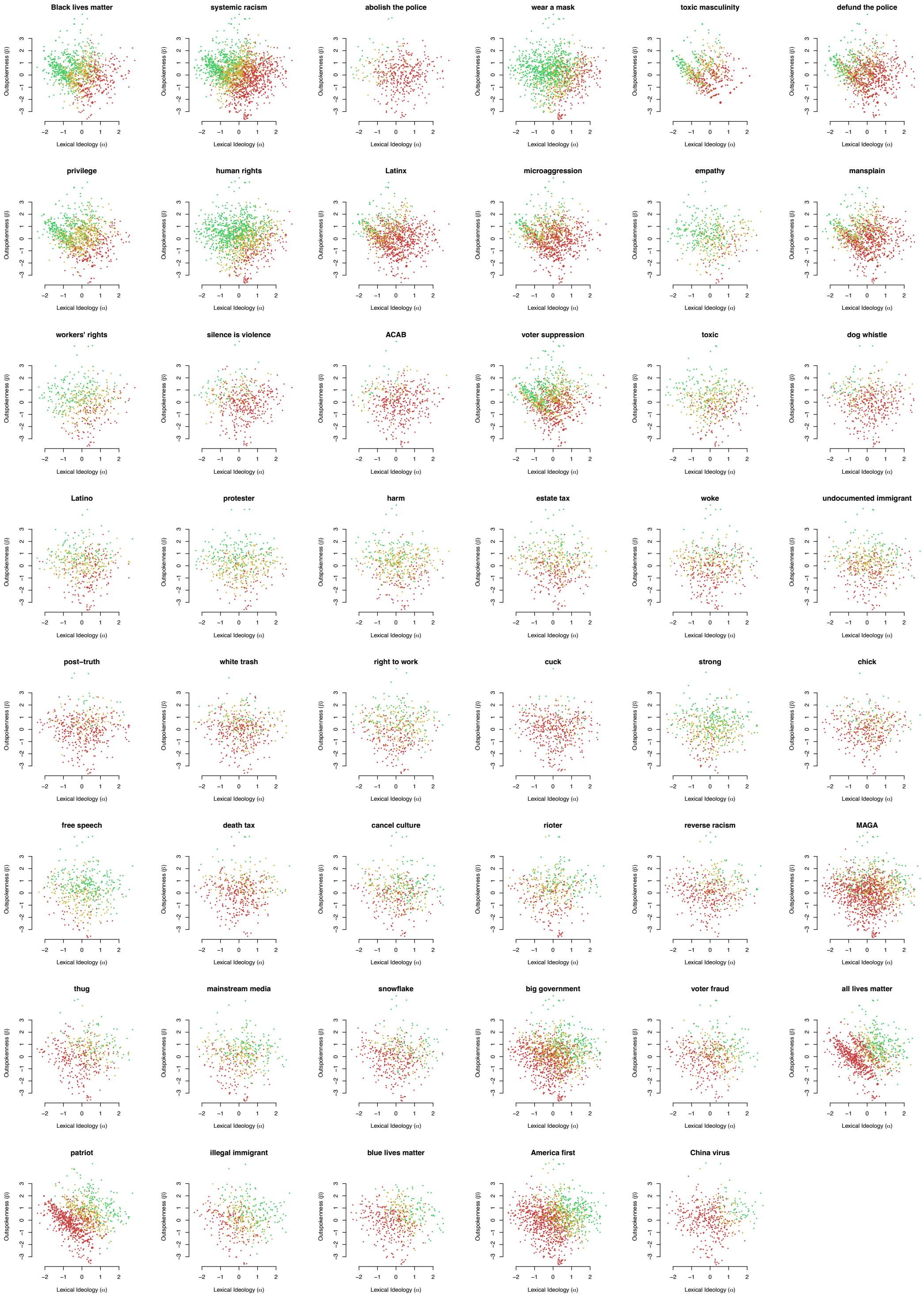


Figure 18. Bivariate response scatterplots (as in Figure 1b) for all 46 phrases measured on a 3-point scale, arranged in order of γ slant. Based on data from Studies 1, 2, and 4, collected 2021-2022.

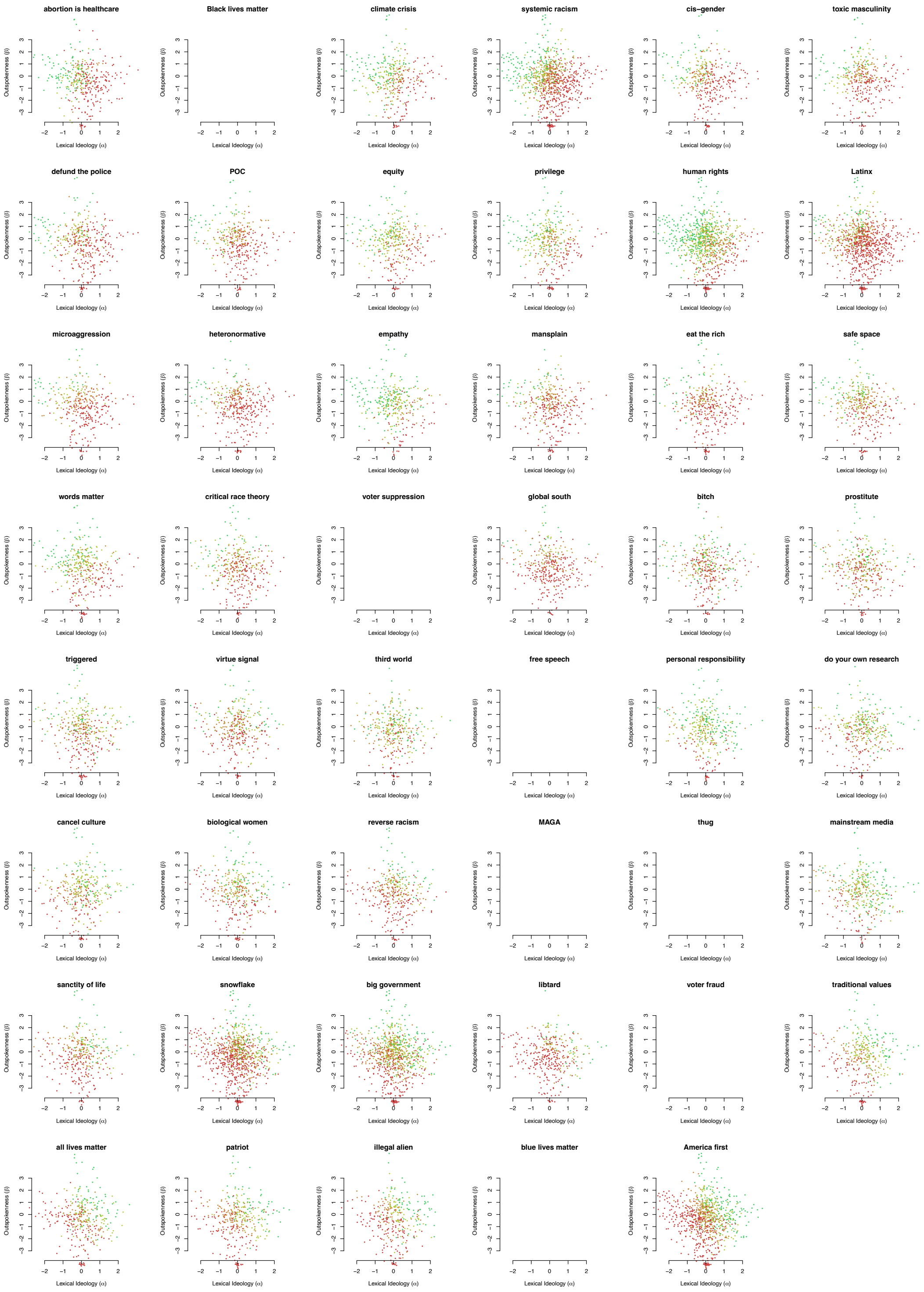


Figure 19. Bivariate response scatterplots (as in Figure 1b) for all 40 phrases measured on a 4-point scale, arranged in order of γ slant. Based on data from Study 4, collected 2023.

H Stick Representations of Phrase Ideologies

I here provide versions of Figure 1c that include all phrases studied on a 3-point scale (Figure 20), and all phrases studied on a 4-point scale (Figure 20). Note that unlike Figure 1c, these plots include some phrases with very weak slants, which are not ideologically-informative (at least in these models). Weakly-slanted phrases are notable for the lack of sharp variation in predicted responses along the lexical ideology spectrum. See, for example, woke in Figure 20, or virtue signal in Figure 21.

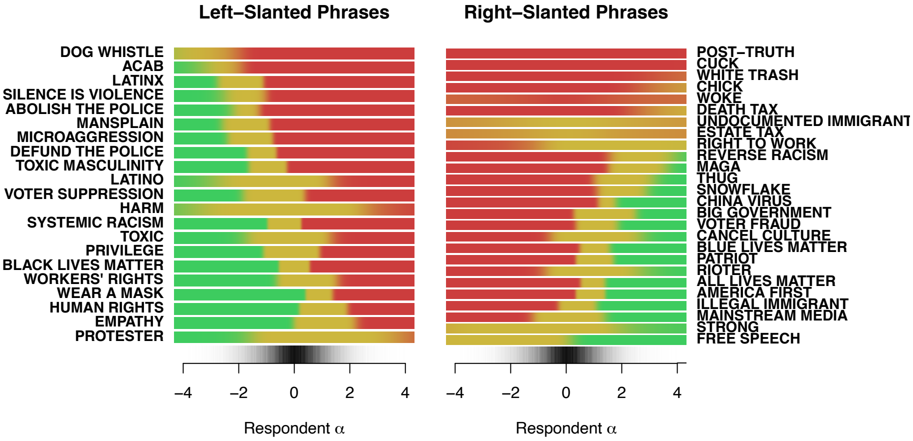


Figure 20. Predicted phrase usage response category as a function of respondent lexical ideology α (as in Figure 1c) for all 46 phrases studied with a 3-point response scale. Based on data from Studies 1, 2, and 4, collected 2021-2022.

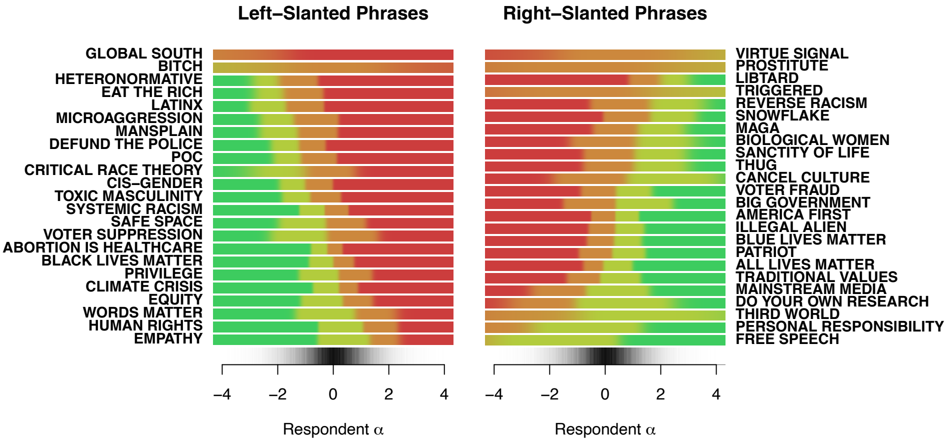


Figure 21. Predicted phrase usage response category as a function of respondent lexical ideology α (as in Figure 1c) for all 40 phrases studied with a 4-point response scale. Based on data from Study 4, collected 2023.

I 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:

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

Study 3 used a slightly different set of issue prompts, in which (2) (7) and (10) in the above list were replaced with the following items:

1. Children should learn in school that the legacy of slavery still affects the position of Black people in American society today.
2. If America is too open to people from all over the world, we risk losing our identity as a nation.
3. People should be free to say whatever they want online, even if others find it offensive or threatening.

J Ideological Identity Strength Questions

The following questions (adapted from Huddy, Mason, and Aarøe 2015) were used in Studies 2 and 4 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

K Validation Study Details

In order to benchmark the WWYS measure against examples of “actual” speech, I conducted a validation exercise using data from a separate research project (conducted in collaboration with Andy Guess) that collected WWYS responses from Twitter users. I then compared the speech trait estimates derived from the WWYS question to characteristics of their actual tweets derived from hand-labeling, in order to verify that the estimates derived from my survey measure are significantly predictive of human-labeled attributes of their online speech.

K.1 Details on Study Furnishing Validation Data

The research project that furnished this validation data was an experiment, conducted with participants who used Twitter, where participants were offered a financial encouragement to follow several left-leaning Twitter accounts. Participants were recruited from Amazon Mechanical Turk via CloudResearch, with filtering criteria set to target US adults with moderate, liberal, or very liberal political views. Additionally, participants were excluded from analysis if: they did not comply with the instructions to follow (and in the case of the control group, subsequently un-follow) a set of 10 left-leaning Twitter accounts, or had twitter accounts that were "protected" (which would prevent data collection), or were created less than 90 days prior to treatment assignment, or followed no other accounts prior to treatment assignment, or did not tweet in the 90 days prior to treatment assignment.

The sampling strategy did not seek to include conservatives due to the nature of the treatment and hypotheses the study was designed for. This generally left-of-center sample creates a relatively stringent validation test for the WWYS method, since the analysis depends on variation in a relatively narrow segment of the ideological spectrum.

From an initially recruited sample of 748 survey respondents, these exclusion criteria reduced the sample down to 270 individuals included in the tweet analysis.

K.2 Special WWYS Question Version

Anticipating the recruitment of a generally left-leaning sample, a special version of the WWYS question was prepared for this study, which used 11 phrases that were expected to be especially useful for capturing variation amongst relatively liberal respondents:

- VOTER SUPPRESSION
- PRIVILEGE
- SYSTEMIC RACISM
- DEFUND THE POLICE
- LATINX
- MANSPLAIN
- TOXIC MASCULINITY
- ALL LIVES MATTER
- MICRO-AGGRESSION
- BLACK LIVES MATTER
- PATRIOT

I developed this list by estimating the Wordsticks model on a subset of the Study 2 respondents who answered the original WWYS question (implemented with the phrases listed in Appendix B) in the hypothetical context of posting on a social media platform. For the purposes of identifying phrases that would be particularly informative in a liberal sample of respondents, I selected phrases for which the estimated γ parameter in this subset model had relatively large absolute values, and which had relatively high variance in the raw data (the ordinal response of “would” (2) “might” (1) or “wouldn’t” (0)), while still maintaining topical breadth. Additionally, I changed TOXIC to TOXIC

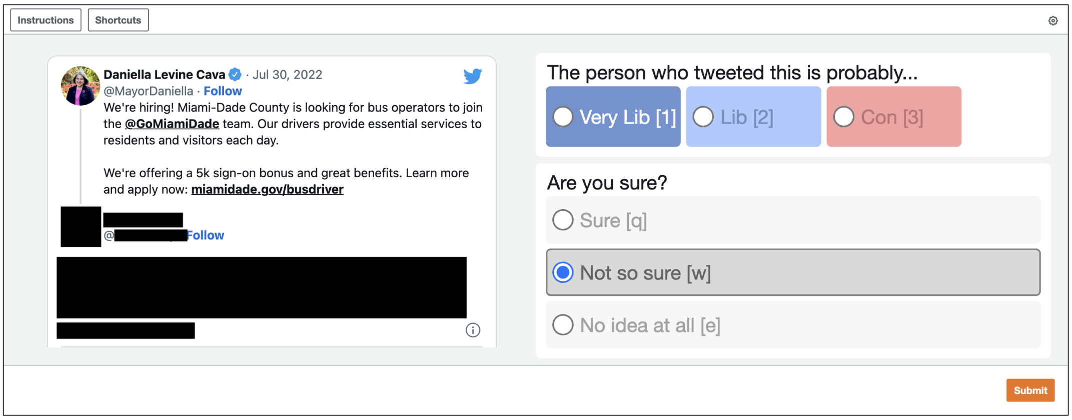
MASCULINITY, purely out of substantive interest.

I implemented the WWYS question in the survey, asking respondents whether they would use each word or phrase "on Twitter," in order to gather WWYS data pertinent to the benchmarking data (tweets).

K.3 Tweet Sampling and Hand-Labeling Procedure

The hand-labeled tweets that constitute the benchmark of "real" speech in this validation exercise were collected and labeled for a separate project, with the purpose of estimating a treatment effect on the ideological content of participants' tweets. Participants' tweets were scraped using the R package `rtweet`, and due to the nature of that study, tweets were sampled for hand-labeling in a particular way: up to 5 tweets per user were sampled in the pre-treatment period from March 16th 2022 through to time of treatment (which varied between participants from mid-June to mid-July 2022), and up to 5 tweets per user were sampled in the post-treatment period from time of treatment through August 19th 2022. The division of the pre-treatment and post-treatment periods are irrelevant for this validation analysis (particularly since the treatment effects on tweets were null); the important implication of this sampling procedure is that the number of tweets sampled per user varied between 0 (if the user did not tweet at all during either time period) and 10 (if the user tweeted at least 5 times during both time periods). In practice, no users had zero tweets, because the prior inclusion criteria for the study required that participants have tweeted in the 90 days prior to treatment assignment. This resulted in a sample of 2,163 tweets (from 270 users) selected for hand-labeling.

Three coders (the author, the author's advisor Andy Guess, and a research assistant) labeled the tweets' political content. To facilitate the labeling process, I implemented a custom annotation interface (see Figure 22) with several key features: first, each tweet was labeled with an ideology of either "Very Liberal," "Liberal," or "Conservative," (two levels of liberalism were included because the sample was generally left-of-center and so it was anticipated that it would be valuable to label finer degrees of variation on the left than on the right), and the annotator was asked to indicate their sureness regarding this ideology label as either "Sure," "Not so sure," or "No idea at all" (in the latter case, no ideology label was required). Annotators viewed the tweets through an embedded iframe that showed tweets in context – for example, if a user's tweet was a reply to another tweet the original tweet was visible for context, as shown in Figure 22. This helped annotators interpret ambiguous tweets that would have been difficult to label in the absence of this context, although it required participants' accounts to be active and public at the time that data was prepared for annotation and at the time of annotation, and that tweets not be deleted before the time of annotation. As a result, 50 tweets were dropped at time of data preparation (from 5 users with 10 tweets each), and 113 tweets were dropped at time of labeling because they could not be viewed by an annotator (two of three annotators labeled tweets as "No idea at all" in these cases, and the third annotator abstained from labeling these cases). As a result, labels were recorded for 2,000 tweets (the roundness of this number was pure coincidence) from 265 users.



Daniella Levine Cava · Jul 30, 2022
@MayorDaniella · Follow
We're hiring! Miami-Dade County is looking for bus operators to join the @GoMiamiDade team. Our drivers provide essential services to residents and visitors each day.

We're offering a 5k sign-on bonus and great benefits. Learn more and apply now: miamidade.gov/busdriver

[redacted]
@[redacted] Follow
[redacted]

The person who tweeted this is probably...
○ Very Lib [1] ○ Lib [2] ○ Con [3]

Are you sure?
○ Sure [q]
● Not so sure [w]
○ No idea at all [e]

Figure 22. Annotation interface

To assess inter-coder agreement, 102 tweets were labeled in triplicate (by all three coders). All three coders' sureness labels were in agreement for 79% of these tweets (which is more than satisfactory, especially considering that all three coders could have chosen from three labels). Table 3 shows agreement on the ideology labels for different subsets of tweets depending on aggregate sureness. Note that sureness was generally low: tweets for which all three coders said they were "sure" constituted 2% of this set of tweets, and tweets that at least two out of three coders said something other than "no idea at all" (the most lenient standard of sureness) constituted 4% of the tweets. As can be seen in Table 3, agreement was very high regarding whether tweets were liberal or conservative, but lower regarding whether tweets were very liberal, liberal, or conservative. This is to be expected, since the distinction between "very liberal" and "liberal" is much less clear than between "liberal" and "conservative." Also, agreement over three categories is necessarily less likely than agreement over two categories, and so it is difficult to make a direct comparison between the two. It should also be noted that, given the low rate of sureness (that is, the low prevalence of identifiably ideological tweets) in this set, these agreement statistics summarize a very small number of tweets, and should not be over-interpreted. To the extent that the hand labels are noisy, the present exercise represents a conservative validation test of the WWYS procedure.

	% Agree (Lib/Con)	% Agree (V. Lib/Lib/Con)	% Batch
All Sure	100	46	2
Most Sure	94	47	2
All not NIAA	89	47	3
Most not NIAA	88	50	4

Table 3. Ideology Agreement, Subset by Sureness

K.4 Validation Analyses

I compared respondents' latent trait estimates to the hand-labeled ideological content of their tweets, to establish the validity of the trait estimates as an estimate of "real" speech.

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 an ideology label to analyze. 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 strongly significant linear relationship ($p<0.01$, see Table 4; Pearson's $\rho = 0.29$, p<0.001). Figure 23 plots the relationship between lexical ideology and mean tweet ideology, shading points according to the number of tweets available from each user. Although many users had only one ideological tweet available for analysis, there is a clearly visible relationship between the lexical ideology trait estimated by the WWYS survey method and the observed ideological slant of users' online speech. This validates the use of self-reported phrase usage as a proxy for the overall slant of individuals' actual speech behavior.

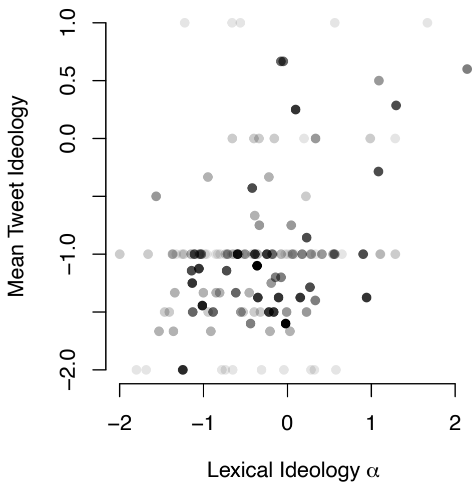


Figure 23. Plot of tweet validation analysis for WWYS α lexical ideology coefficient. Points represent users, and are shaded according to the number of tweets available from that user, ranging from 1 tweet (light grey) to 10 tweets (black). Based on data from Study 4, collected 2022.

Table 4. Alpha Tweet Validation

	mean_ideo3_conf_crossed
alpha	0.49
	(0.13)
Constant	−1.57
	(0.11)
Observations	148
Adjusted R^2	0.08

I also tested whether the WWYS outspokenness estimates corresponded to the hand-labeled "sureness" that coders reported regarding the ideological coding of users' tweets. To do this, 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 is much noisier than in the ideology analysis, but still statistically significant (p<.05, see Table 5; Pearson's ρ=0.16, p<0.05). This was not the intended purpose of measuring sureness, however it serves as a supplementary validation of the interpretation of the β trait as “outspokenness,” since the more politically-outspoken an individual is, we should expect them to express their political opinions more frequently and clearly, such that an observer can more confidently infer their political position from what they say.

Table 5. Beta Tweet Validation

	mean_conf
beta	0.08
	(0.03)
Constant	0.45
	(0.04)
Observations	231
Adjusted R^2	0.02

L Beta News Regressions

Table 6. Beta Regressions: No Media and Disaggregated Media

	No News	News Disaggregated
	(1)	(2)
age_dec	0.03 (0.03)	0.03 (0.03)
ideo_6	0.02 (0.06)	−0.01 (0.06)
PID_6	−0.08 (0.06)	−0.06 (0.06)
pol_int	0.27 (0.04)	0.14 (0.04)
identity_strength	0.13 (0.02)	0.09 (0.02)
POC	0.10 (0.09)	0.10 (0.09)
male	−0.09 (0.07)	−0.08 (0.07)
treatmentliberal	−0.07 (0.11)	−0.10 (0.11)
treatmentconservative	−0.24 (0.11)	−0.23 (0.11)
treatmentsm	−0.21 (0.10)	−0.24 (0.10)
abs(issue_scale)	−0.28 (0.06)	−0.16 (0.06)
as.factor(study)3	−0.27 (0.10)	−0.28 (0.09)
tv		0.01 (0.02)
newspapers		0.07 (0.03)
radio		0.11 (0.03)
internet_sm		−0.01 (0.02)
discussions		0.14 (0.03)
podcasts		0.08 (0.03)
Constant	0.24 (0.15)	−0.49 (0.18)
Observations	1,698	1,698
Adjusted R^2	0.10	0.15

M Study 3 Regression Details and Power Analyses

This appendix presents supplementary information on the analyses of the experimental treatment in Study 3, which asked participants to imagine speaking with the most liberal and conservative people they talk to (in addition to the close-friend control condition).

First, Tables 7 and 8 present the regression analyses used to produce Figure 5 in the main text.

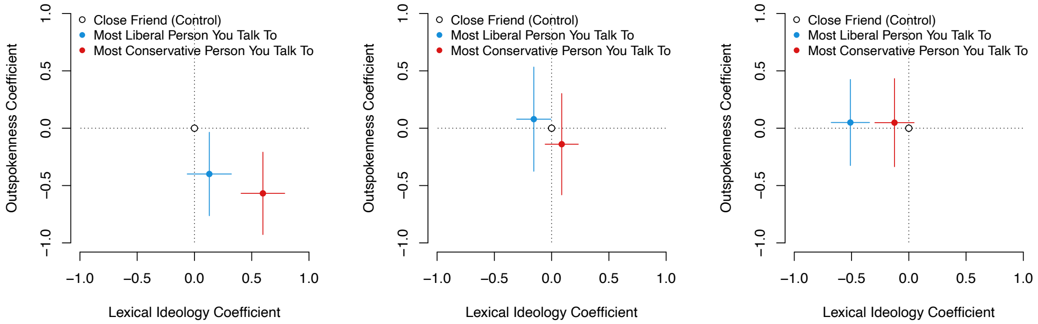
Table 7. Study 3 Subgroup Regressions (Alpha, Pooled Study 1,3 Data)

	Liberals	Moderates	Conservatives
	(1)	(2)	(3)
treatmentstranger	0.05 (0.11)	−0.05 (0.10)	−0.12 (0.11)
treatmentliberal	0.02 (0.10)	−0.14 (0.09)	−0.53 (0.10)
treatmentconservative	0.62 (0.10)	0.11 (0.09)	−0.15 (0.10)
identity_scale	0.20 (0.10)	0.10 (0.09)	0.36 (0.09)
ideo_6	0.70 (0.12)	0.28 (0.18)	0.62 (0.14)
age_dec	0.11 (0.03)	0.06 (0.03)	0.07 (0.02)
college	−0.03 (0.07)	−0.04 (0.06)	−0.08 (0.07)
male	0.27 (0.07)	0.19 (0.06)	0.08 (0.07)
POC	0.11 (0.09)	−0.13 (0.07)	−0.30 (0.10)
as.factor(study)3	−0.003 (0.11)	−0.04 (0.10)	−0.03 (0.10)
Constant	−0.62 (0.19)	−0.31 (0.14)	−0.61 (0.22)
Observations	515	400	489
Adjusted R^2	0.23	0.10	0.21

Table 8. Study 3 Subgroup Regressions (Beta, Pooled Study 1,3 Data)

	Liberals	Moderates	Conservatives
	(1)	(2)	(3)
age_dec	−0.07 (0.05)	−0.02 (0.06)	0.11 (0.05)
ideo_6	0.06 (0.20)	0.11 (0.25)	−0.22 (0.27)
PID_6	0.32 (0.12)	−0.14 (0.14)	−0.13 (0.10)
pol_int	0.01 (0.07)	0.19 (0.08)	0.01 (0.08)
identity_strength	0.18 (0.05)	0.13 (0.07)	0.14 (0.05)
POC	0.04 (0.15)	−0.10 (0.17)	0.36 (0.19)
male	−0.30 (0.11)	−0.18 (0.15)	−0.08 (0.13)
news_scale	0.36 (0.05)	0.22 (0.07)	0.35 (0.06)
treatmentstranger	−0.47 (0.18)	−0.60 (0.23)	−0.64 (0.20)
treatmentliberal	−0.42 (0.17)	0.07 (0.21)	0.03 (0.18)
treatmentconservative	−0.58 (0.17)	−0.14 (0.21)	0.04 (0.18)
as.factor(study)3	−0.44 (0.17)	−0.52 (0.22)	−0.85 (0.19)
Constant	1.11 (0.33)	0.62 (0.35)	0.47 (0.45)
Observations	514	399	488
Adjusted R^2	0.21	0.10	0.17

Next, Figure 24 provides a version of Figure 5 where regression models (see Tables 9 and 10) were estimated solely on the Study 3 data. Substantively, these results are essentially identical to those reported in the main text. The only difference is that these models lose the information about the close-friend treatment (that was also used in Study 1), and so the estimates of the β treatment effects lose some precision; however these models actually gain precision in estimating α, because they are able to condition on issue ideology (which was not measured in Study 1 and so could not be included in the models in Tables 7 and 8 that pooled the Study 1 and 3 data).



(a) Liberal Respondents

(b) Moderate Respondents

(c) Conservative Respondents

Figure 24. Study 3 treatment coefficients (and 95% CIs) subset by respondent self-described ideology: Liberals (Panel a, see Model 1 of Tables 9 and 10), Moderates (Panel b, see Model 2 of Tables 9 and 10), and Conservatives (Panel c, see Model 3 of Tables 9 and 10). Data from Study 3, collected 2023.

Table 9. Study 3 Subgroup Regressions (Alpha, Study 3 Data Only)

	Liberals	Moderates	Conservatives
	(1)	(2)	(3)
treatmentliberal	0.13 (0.10)	−0.16 (0.08)	−0.51 (0.09)
treatmentconservative	0.60 (0.10)	0.09 (0.07)	−0.12 (0.09)
issue_scale	0.61 (0.07)	0.32 (0.04)	0.45 (0.05)
identity_scale	0.23 (0.12)	−0.04 (0.09)	0.24 (0.10)
ideo_6	0.42 (0.15)	0.21 (0.17)	0.43 (0.16)
age_dec	0.11 (0.03)	0.04 (0.03)	0.07 (0.03)
college	0.05 (0.08)	−0.02 (0.06)	−0.05 (0.07)
male	0.14 (0.08)	0.14 (0.07)	−0.02 (0.07)
POC	0.01 (0.11)	−0.08 (0.07)	−0.02 (0.11)
Constant	−0.35 (0.21)	−0.19 (0.14)	−0.59 (0.22)
Observations	325	256	320
Adjusted R^2	0.38	0.28	0.40

Table 10. Study 3 Subgroup Regressions (Beta, Study 3 Data Only)

	Liberals	Moderates	Conservatives
	(1)	(2)	(3)
age_dec	−0.06 (0.06)	−0.03 (0.09)	0.17 (0.06)
ideo_6	0.18 (0.28)	0.11 (0.34)	−0.19 (0.36)
PID_6	0.27 (0.14)	−0.13 (0.19)	−0.16 (0.13)
pol_int	0.01 (0.10)	0.26 (0.11)	0.07 (0.10)
identity_strength	0.19 (0.07)	0.12 (0.09)	0.12 (0.07)
POC	0.04 (0.21)	−0.13 (0.22)	0.09 (0.26)
male	−0.40 (0.16)	−0.22 (0.20)	0.02 (0.17)
news_scale	0.39 (0.07)	0.22 (0.09)	0.29 (0.08)
treatmentliberal	−0.40 (0.19)	0.08 (0.23)	0.05 (0.19)
treatmentconservative	−0.57 (0.18)	−0.14 (0.23)	0.05 (0.20)
abs(issue_scale)	−0.09 (0.18)	−0.03 (0.21)	−0.28 (0.13)
Constant	0.82 (0.42)	0.20 (0.49)	−0.41 (0.55)
Observations	324	256	319
Adjusted R^2	0.21	0.07	0.12

M.1 Power Analyses

Finally, I conduct power analyses to inform the interpretation of the Study 3 results presented above. Because previous literature offers little guidance as to the plausible size of these treatment effects, I conduct power analyses by simulation: I ran 10,000 simulations in which I sampled observations (with replacement) from the Study 3 dataset[6] (holding the total N fixed), re-estimated the regressions presented in Tables 9 and 10, and calculated the proportion of simulations where each treatment (in each ideological subset of the respondents) was estimated to have a treatment effect that was statistically significant at the conventional .05 level. This provides an estimate of Study 3's power to detect treatment effects with conventional levels of significance, assuming that the point estimates found in Study 3 (see Tables 9 and 10) are reasonable approximations of the true effect size.

Table 11 provides estimates of Study 3's power to detect treatment effects on α, for each treatment (Most Liberal Person You Talk To, Most Conservative Person You Talk To) relative to the Close-Friend control condition. Study 3 appears to have had excellent power to detect the "accomodation" effects of the Conservative treatment on Liberals, and the Liberal treatment on Conservatives (none of 10,000 simulations failed to find these effects significant at the .05 level).

If we take the Study 3 point estimates of the "in-group" treatment effects (of the Liberal treatment on Liberals, and the Conservative treatment on Conservatives) as a reasonable approximation of their true magnitude, Study 3 was not well-powered to detect these effects. However, the estimates of these treatment effects have the opposite sign from what one would predict based on social identity theory: we would expect liberals and conservatives to conform to their group-prototypes under these treatments, and adopt more liberal and conservative lexica, respectively. So, this power analysis does not change the substantive interpretation of no polarization under the "in-group" treatments treatments. Moreover, social identity theory would predict that individuals conform *more* to their in-group than their out-group – if this had been the case, Study 3 would have had excellent power to detect these effects, as discussed above.

Study 3 was not very well powered to detect treatment effects amongst moderate respondents (53% power for the Liberal treatment, 23% power for the Conservative treatment). The signs of these effects are consistent with conformity, and (notably) the Liberal treatment effect on Moderates reaches conventional levels of statistical significance in the version of the analysis that uses only the Study 3 data (see Table 9, Model 2). So, there is reason to believe that the Liberal treatment shifted Moderates' political lexica meaningfully leftward, and due to the limited power of the Study 3 analysis of Moderates, we cannot rule out a corresponding rightward shift in Moderates' lexica under the Conservative treatment. So, it is plausible that Moderates conform to ideologically-extreme interlocutors, but further research would be needed to reach firm conclusions.

	Treatment	Subset	Power
1	Liberal Interlocutor	Liberals	0.24
2	Conservative Interlocutor	Liberals	1.00
3	Liberal Interlocutor	Conservatives	1.00
4	Conservative Interlocutor	Conservatives	0.29
5	Liberal Interlocutor	Moderates	0.53
6	Conservative Interlocutor	Moderates	0.23

Table 11. Study 3 Simulated Power (Alpha)

[6] Note that I used the Study 3 data only, rather than the pooled Study 1 and Study 3 data. As noted above, using the Study 1 data does not affect the substantive results for the Study 3 treatment effects, and the power analysis is more interpretable when focused on the Study 3 data alone.

Although outspokenness was not the main outcome of interest in Study 3, Table 12 provides corresponding power estimates for completeness. The study appears to have been fairly well-powered to detect the β effects detected amongst liberals in both treatments, but was not well-powered to detect effects on β for Liberals and Moderates.

	Treatment	Subset	Power
1	Liberal Interlocutor	Liberals	0.58
2	Conservative Interlocutor	Liberals	0.86
3	Liberal Interlocutor	Conservatives	0.04
4	Conservative Interlocutor	Conservatives	0.04
5	Liberal Interlocutor	Moderates	0.05
6	Conservative Interlocutor	Moderates	0.10

Table 12. Study 3 Simulated Power (Beta)

N Study-Specific Experiment Regressions

Table 13. Alpha Regressions (Studies 1 and 3 Separately)

	Study 1	Study 3
	(1)	(2)
t=Stranger	−0.04 (0.06)	
t=Liberal		−0.20 (0.06)
t=Conservative		0.22 (0.06)
Age (Decades)	0.11 (0.02)	0.08 (0.02)
6-Point Ideo	0.49 (0.06)	0.50 (0.04)
6-Point PID	0.28 (0.06)	0.14 (0.04)
College	−0.14 (0.06)	0.01 (0.05)
Male	0.21 (0.06)	0.21 (0.05)
POC	0.03 (0.08)	−0.09 (0.07)
Intercept	−0.59 (0.12)	−0.56 (0.10)
Observations	502	899
Adjusted R^2	0.57	0.48

Table 14. Beta Regressions (Studies 1 and 3 Separately)

	Study 1	Study 3
	(1)	(2)
Age (Decades)	−0.02 (0.04)	0.01 (0.04)
6-Point Ideo	−0.01 (0.10)	0.04 (0.08)
6-Point PID	0.04 (0.10)	−0.03 (0.08)
Political Interest	0.01 (0.06)	0.08 (0.06)
Identity Strength	0.18 (0.04)	0.09 (0.03)
POC	0.18 (0.14)	0.06 (0.13)
Male	−0.11 (0.11)	−0.25 (0.10)
News Scale	0.29 (0.05)	0.35 (0.04)
t=Stranger	−0.58 (0.10)	
t=Liberal		−0.12 (0.12)
t=Conservative		−0.23 (0.11)
Intercept	0.50 (0.20)	−0.11 (0.20)
Observations	502	899
Adjusted R^2	0.19	0.13