Political Catchphrases Reveal Polarizing Self-Censorship on Facebook and Twitter

William Small Schulz

Abstract

It is widely agreed that online political discourse is polarized, in that extreme views appear more prevalent online than off. I conduct a pre-registered experiment with a representative sample of Facebook and Twitter users, to test two mechanisms theorized to drive this polarization: preference falsification (users exaggerating their extremity online) and self-censorship (moderates refraining from speaking up online). Using an original method that exploits contemporary political catchphrases to summarize survey respondents' political expression, I estimate differences in how people speak about politics online versus offline, and measure the ideological distribution of the speech that is "missing" from platforms due to users' self-censorship. Results favor self-censorship over preference falsification: users who talk about politics online have more polarized speech than those who stay silent, but there is little evidence that users adopt more polarized speech online than offline. This suggests that depolarizing online discourse and enhancing users' freedom-of-speech are compatible goals.

Keywords: Speech, Polarization, Social Media, Self-Censorship, Preference Falsification

Gregory	Is there any point to which you would wish to draw my attention?
Holmes	To the curious incident of the dog in the night-time.
Gregory	But the dog did nothing in the night-time.
Holmes	*That* was the curious incident.

Conan Doyle, "The Adventure of Silver Blaze" (1892)

Most Americans perceive online political discourse to be excessively polarized (Gallup 2022). At the same time, most Americans hold political views that are best described as moderate, or at least nuanced (Fowler et al. 2023). Together, these facts suggest that online platforms fail to accurately represent the distribution of US public opinion, but the exact mechanism of this misrepresentation is disputed.

One popular theory is that social media users engage in a form of *preference falsification*: that they experience social pressure to feign attitudes (Kuran 1995, see also Goffman 1956) more extreme than their true beliefs, or the beliefs that they would express in traditional offline contexts. In many ways, this seems a natural inference to draw from the apparent disjuncture between online and offline discourse, and several prominent scholars have advanced theories along these lines. For example, Sunstein (2017) has theorized that like-minded social influence engenders "reputational cascades" (p. 102) in which users conform to their political tribes, creating a discourse-polarizing feedback loop. Haidt (2022) similarly argues that platforms' publicity, quantification of likes, and algorithmic amplification have created a new kind of social game that "encourage[s] dishonesty and mob dynamics: users [are] guided not just by their true preferences but by their past experiences of reward," and thereby pollute online discourse with performative partisanship.

Perhaps because this theory aligns with contemporary concerns about "virtue signaling," (Hill and Fanciullo 2023), it has wide bipartisan appeal: 97% of Democrats and 94% Republicans believe that platforms induce or enable people to say things that they would not say in a face-to-face conversation (Gallup 2022). However, there is strikingly little evidence to support this view, because of a simple but stubborn problem: although we can observe social media users' online speech, it is much harder to observe their face-to-face conversations, making it difficult to estimate alleged differences in self-expression between online and offline settings. As a result, theories of polarizing online preference falsification depend heavily on data collected in quite different domains, such as laboratory-based studies of group discussions (e.g. McGarty et al. 1992), which may not generalize to contemporary online settings.

Moreover, recent evidence suggests an entirely different mechanism of online polarization: *self-censorship*, which occurs when individuals refrain from expressing their true preferences, for fear of social backlash (see Noelle-Neumann 1974). According to this view, the apparent polarization of online discourse arises from the tendency of polarized individuals to self-select into sharing their views online, while more staid users remain silent, leading to the over-representation of polarized speech on social platforms. This builds on evidence that so-called "lurkers" make up the majority of most online communities (Nonnecke and Preece 2000), such that most political content comes from a small minority of users who have unrepresentatively extreme views (Wojcik and Hughes 2019).

The self-censorship theory is corroborated by several recent studies. Bor & Petersen's (2022) investigation of online *hostility* found that users were no more hostile online than off, but non-hostile individuals tended to abstain from online political discussion, producing "adverse selection bias" (p. 1) in favor of hostility. Similarly, Kim et al. (2021) found that the *toxicity* of online comment threads was exacerbated by the tendency of attitudinally-polarized and toxicity-prone users to select into commenting in the first place. Based on these findings, it is plausible that the apparent *polarization* of online discourse may arise not from individuals exaggerating their polarization online compared to offline, but instead from this same form of adverse self-selection of online speakers that under-represents moderates. Indeed, Bail (2021) describes moderates' self-censorship as "the most profound form of distortion created by the social media prism." (p. 82)

However, if this polarizing distortion is caused by individuals self-selecting out of online discourse, no "big data" analysis of social media traces can reveal it, since it necessarily selects on the dependent variable of interest: online speech. To know the effects of self-censorship, we require some estimate of what a person *would have said* if they *hadn't* self-censored.

At present, the preference falsification theory commands wide credence despite a narrow evidence base, while recent studies support self-censorship without conclusively ruling out falsification. Clearer evidence is needed, especially given the importance of polarization as a social problem, and the fact that preference falsification and self-censorship have quite different practical implications for the improvement of online discourse (a topic I return to in the Discussion). In this study, therefore, I implement an original experiment[1] that addresses the methodological challenges outlined above, in order to test *both* theories within a shared framework.

METHOD

I apply an original method (Anonymized) to estimate differences between social media users' online and offline political expression, through a unique experiment employing the specialized survey instrument shown in Figure 1. This "What Would You Say?" (WWYS) question asks respondents whether they would use politically-charged catchphrases, like "systemic racism" and "big government," in a given context, such as posting online or talking with a friend.

Because these phrases signify ideological positions (see Figure 2, discussed further in Results), I can scale self-reported phrase usage using an ordinal version of Slapin and Proksch's (2008) Wordfish model, to estimate an ideal point (a "lexical ideology") for each respondent, as well as each respondent's propensity to use such phrases at all (their "outspokenness").

Because the context specified in the WWYS question can be manipulated in a between-subjects experiment, I am able to estimate causal effects of context on these two dimensions of speech. This furnishes a pragmatic operationalization of preference falsification as a kind of ideological code-switching: do people talk "more liberally" or "more conservatively" online compared to offline? I am able to test for polarizing preference falsification by analyzing these context-driven shifts in lexical ideology.

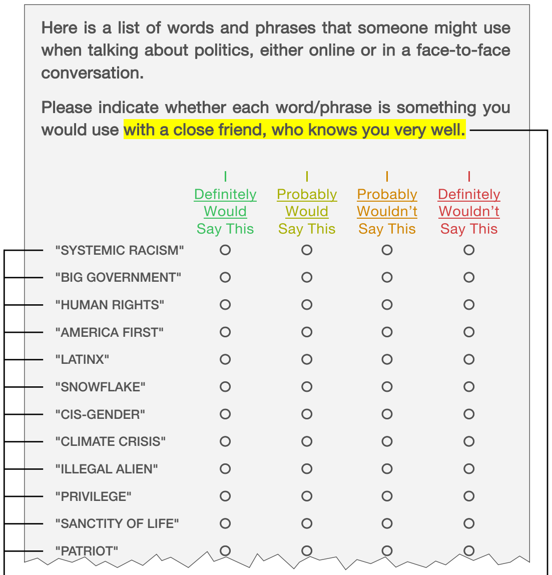
And, because the WWYS question can be posed to a representative sample of social media users, including those who avoid talking about politics online, it can measure the speech that is *missing* from online platforms, by using self-censorers' offline speech as a proxy. This permits a test of polarization by self-censorship.

[1] The experimental procedure and hypothesis tests were pre-registered. Anonymized pre-registration documents accompany this manuscript.

FIGURE 1. The “What Would You Say?” Question, Context Treatments, and Phrases

WWYS QUESTION

The "What Would You Say?" (WWYS) question asks whether the respondent would use certain words and phrases, in a specified social context:



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

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

	Definitely Would Say This	Probably Would Say This	Probably Wouldn't Say This	Definitely Wouldn't Say This
"SYSTEMIC RACISM"	○	○	○	○
"BIG GOVERNMENT"	○	○	○	○
"HUMAN RIGHTS"	○	○	○	○
"AMERICA FIRST"	○	○	○	○
"LATINX"	○	○	○	○
"SNOWFLAKE"	○	○	○	○
"CIS-GENDER"	○	○	○	○
"CLIMATE CRISIS"	○	○	○	○
"ILLEGAL ALIEN"	○	○	○	○
"PRIVILEGE"	○	○	○	○
"SANCTITY OF LIFE"	○	○	○	○
"PATRIOT"	○	○	○	○

CONTEXT TREATMENTS

The context specified in the WWYS question can be used as an experimental treatment. In this case, it is manipulated to refer to an offline or an online context:

with a close friend, who knows you very well.

on Twitter.

on Facebook.

PHRASES

Respondents report their hypothetical usage of 20 phrases. The first 6 phrases are held constant, and the remaining 14 are sampled from a larger superset. In total, the study asks about usage of 40 phrases:

SYSTEMIC RACISM BIG GOVERNMENT HUMAN RIGHTS
AMERICA FIRST LATINX SNOWFLAKE CIS-GENDER
BLACK LIVES MATTER ABORTION IS HEALTHCARE
HETERONORMATIVE PRIVILEGE TOXIC MASCULINITY
CLIMATE CRISIS EQUITY EMPATHY MICRO-AGGRESSION
DEFUND THE POLICE SAFE SPACE POC MANSPLAIN
WORDS MATTER EAT THE RICH VOTER SUPPRESSION
ALL LIVES MATTER LIBTARD SANCTITY OF LIFE
REVERSE RACISM VOTER FRAUD ILLEGAL ALIEN
TRADITIONAL VALUES FREE SPEECH BLUE LIVES MATTER
MAGA MAINSTREAM MEDIA DO YOUR OWN RESEARCH
THUG CANCEL CULTURE PATRIOT BIOLOGICAL WOMEN
PERSONAL RESPONSIBILITY

So, the WWYS method solves the key methodological challenges identified above: it can measure differences in speech between contexts, and it can characterize the offline speech of self-censorers. Prior work (Anonymized) has demonstrated the WWYS measure, and validated it against hand-labeled social media posts collected from survey respondents. Thus it provides an ideal tool for the present investigation.

PROCEDURE

To test the two polarization mechanisms, I designed a survey experiment using the WWYS method. To test for polarization by self-censorship, I estimated descriptive differences between the offline speech patterns of “posters” (users who post their political views online) and “lurkers” (users who abstain from posting their political views). To test for polarization by preference falsification, I estimated causal differences between posters’ online and offline speech induced by the context treatments (Twitter/Facebook *vs* “close friend”).

I fielded this experiment in a large representative sample of Facebook and Twitter users drawn from the AmeriSpeak panel maintained by NORC at the University of Chicago (see Table 1 for sample sizes). My questionnaire (see Appendix F) divided respondents into posters and lurkers based on the following questions:

1. Whether or not they used each of 10 online platforms, including Twitter and Facebook.
2. Which of these platforms they used “to post your opinions about politics or current events.”

Participants were eligible as “Facebook posters” if they selected Facebook in both questions; if they selected Twitter in both questions they could be considered "Twitter posters." If a participant selected a platform in the first question, but not the second, they were eligible to be considered a "lurker" on that platform. Participants who used neither Facebook nor Twitter were ineligible for further participation and exited the survey, and any who qualified for multiple groups were assigned to the least-filled group at time of recruitment.[2]

Next, participants answered the WWYS question, which included an experimental manipulation for posters: I randomized the WWYS question to ask which phrases the respondent would use either "with a close friend, who knows you very well," or "on Twitter" (for Twitter-posters) or "on Facebook" (for Facebook-posters). This permits estimation of preference falsification as a causal effect of platform context – a "platform effect" – relative to speaking with a close friend (which is a meaningful alternative to participating in online political discourse, and theoretically elicits a relatively authentic mode of self-presentation, making it a useful reference point for measuring online falsification).

Lurkers, meanwhile, always received the close-friend condition. It wouldn't make sense to ask about their (nonexistent) online political speech, but measuring lurkers' close-friend speech permitted a descriptive comparison against posters' close-friend speech. This allowed me to test for polarization by self-censorship, by testing whether posters' *offline* speech is more polarized than lurkers'. If so, this would indicate that the speech that is *missing* from online platforms is systematically more moderate than the speech that occurs.

Before fielding the survey, I pre-registered four hypotheses, which are listed below:

H1 **Among posters, the Twitter/Facebook treatment has a negative effect (relative to the "close friend" condition) on outspokenness.** That is, I predict that posters are less outspoken online than with close friends. This reflects my expectation that users generally self-censor political language from their online speech, relative to how they speak with close friends.

H2 **Among posters, the Twitter/Facebook treatment has a null effect (relative to the "close friend" condition) on lexical ideology.** Rejecting this null hypothesis would indicate that platforms shift posters' speech leftward or rightward, relative to how they speak with close friends. Although such a shift is plausible, I predict a null effect because I have no a priori theoretical reason to expect a shift in a particular direction.

H3 **Posters' close-friend lexical ideology is more polarized (that is, has greater variance) than lurkers' close-friend lexical ideology.** This reflects a self-selection theory of online discourse polarization, in that the people who post their political views online tend to have more polarized speech patterns than the people who don't, as measured from their speech in the close-friend context (which is the context in which posters' and lurkers' speech can be compared).

H4 **Posters' online lexical ideology is *not* more polarized (that is, does *not* have greater variance) than posters' close-friend lexical ideology.** If posters' online lexical

[2] This was to meet quota targets (see Appendix B).

ideology *were* more polarized than posters' close-friend lexical ideology, this would indicate that platforms cause posters to use more polarized political language online than they use offline with their close friends, consistent with a code-switching or preference falsification theory of online discourse polarization. However, I expect that this does not describe most posters' behavior.

These predictions were based on a general expectation that social media users fear being criticized for their political views, and so generally self-censor politics from their online posts (H1). I predicted a null left-right preference falsification effect (H2), absent any theoretical reason to expect an effect in a particular direction. Most importantly, I predicted that online discourse is polarized by moderates' self-censorship (H3), and *not* by posters' preference falsification (H4). Although the latter theory is more popular, the former is better supported by existing evidence, and it is also theoretically *easier* to avoid criticism by simply doing nothing than by feigning extremism.

In order to maximize statistical power, I registered my hypotheses with respect to the pooled Facebook and Twitter respondent data, and reserved platform-specific estimates for exploratory analyses. Also, because Hypotheses 1, 3, and 4 were directional, I pre-registered one-sided tests for these hypotheses, and a two-sided test for Hypothesis 2.

TABLE 1. Recruitment: Quota Targets and Actual Completes

Group	Target	Actual
Facebook poster	1000	1010
Facebook lurker	170	175
Twitter poster	1000	1018
Twitter lurker	170	175

RESULTS

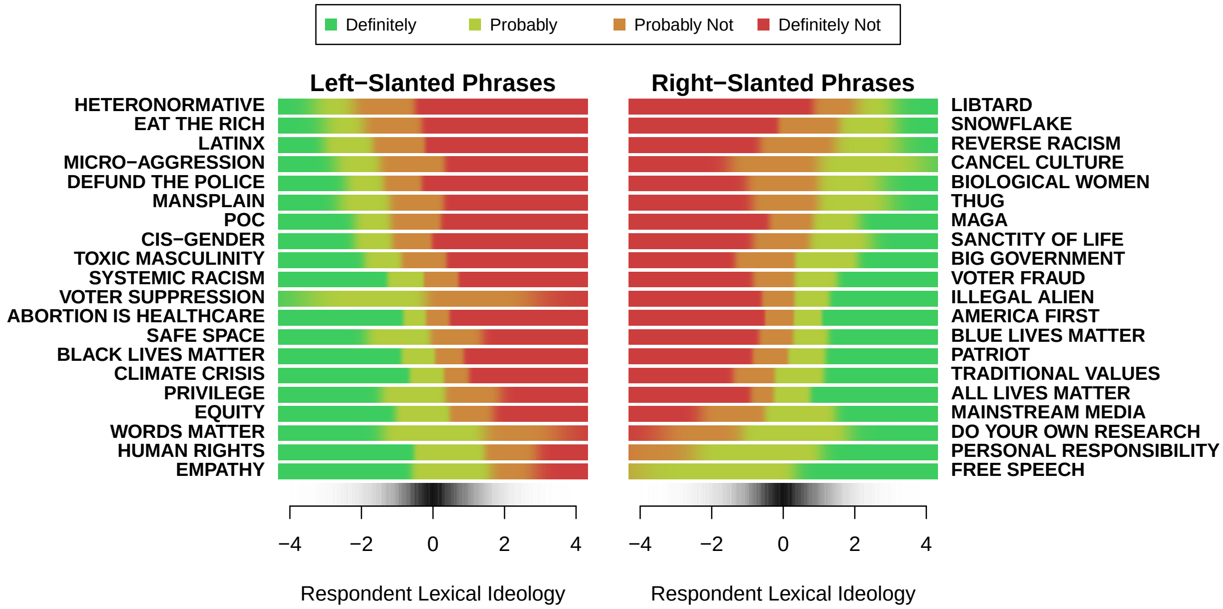
I submitted my pre-registration documents on July 9th, 2023, and NORC collected data from July 11th until August 25th. Table 1 summarizes quota targets and actual completes, by platform and lurker/poster categorization. These targets were based on power-analyses-by-simulation, using data from a pilot (see Appendix C). The only deviation was that self-reported ideology was measured on a 5-point rather than a 7-point scale (which was unavailable).

Visualizing Phrase Ideology

Before presenting the results of the pre-registered hypothesis tests, it is instructive to visualize the ideological content of the catchphrases used to estimate respondents' lexical ideology. Figure 2 therefore plots the ideological positions embodied by each phrase, in terms of the survey responses that would be predicted for respondents at different points along the lexical ideology spectrum. This defines the ideology of each phrase in terms of the ideology of the person who would use it. By this measure, HETERONORMATIVE and LIBTARD are the most extreme left- and right-slanted phrases, respectively, while EMPATHY and FREE SPEECH are the least extreme.

Although the ideological positions embodied by these phrases are not my main object of inquiry, they help contextualize the polarization analyses that I present below, by providing what Monroe et al. (2008) call "semantic validity" (p. 373, see also Krippendorff 2004). In particu-

FIGURE 2. Phrase Ideologies as Predicted Response Regions



Note: Each phrase's ideological position is visualized by plotting the response (color) that would be predicted from respondents across the lexical ideology spectrum. Rugplots on x axis (marked in standard deviations) give the distribution of actual respondents' lexical ideologies, which is $\mathcal{N}(0,1)$ by construction of the model.

lar, they clarify the substantive content of the WWYS measure: more "extreme" speakers are those who are willing to say the most extreme left- or right-slanted phrases, to the exclusion of opposite-slanted phrases. More "outspoken" ones, meanwhile, are those who are more willing to say these kinds of phrases in general, holding their overall ideological mixture fixed.

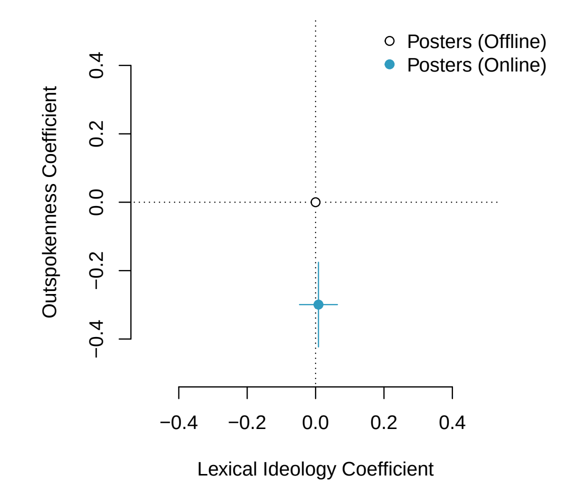
Pre-Registered Analyses

The pre-registered analyses gave results consistent with expectations for all four hypotheses: platforms decreased posters' outspokenness (H1) and had neither a linear (H2) nor a polarizing (H4) effect on their lexical ideology relative to the close-friend condition, but descriptively, posters were found to have significantly more polarized close-friend lexical ideology than lurkers (H3).

These results are visualized in Figures 3-5: Figure 3 plots treatment coefficients (in both the lexical ideology and outspokenness dimensions), from linear regression analyses that were planned to test Hypotheses 1 and 2; meanwhile Figures 4 and 5 plot smoothed lexical ideology densities (bandwidth = .25) to illustrate the variance comparisons that were planned to test Hypotheses 3 and 4, respectively.

Linear Regression (H1 & H2) Pre-registered analyses for Hypotheses 1 and 2 employed linear regression, with Eicker-Huber-White HC2 robust standard errors, implemented in the `estimatr` R package (Blair et al. 2024). Figure 3 plots the pooled platform effects as a linear regression treatment coefficient in two dimensions: outspokenness on the y axis, and lexical ideology on the x axis (see "Pooled" model in Tables 2 and

FIGURE 3. Linear regression platform treatment effects on lexical ideology (H1, X axis) and outspokenness (H2, Y axis).



Note: Posters (Offline) is the reference condition

3, respectively). Compared to the close-friend reference point, the platform treatment significantly decreased outspokenness (p < .00001, pre-registered one-sided test), and had a null effect on lexical ideology (p ≈ .38, pre-registered one-sided test), pre-registered two-sided test). These results were consistent with pre-registered expectations for H1 and H2, respectively.

Exploratory analyses confirmed that the pooled H1 and H2 findings replicated for both Facebook and Twitter users, when analyzed separately (see "Facebook" and "Twitter" models in Tables 2 and 3). So, it appears that users generally avoid political language on both Facebook and Twitter, relative to how they speak with close friends – which is also consistent with recent evidence from Carlson & Settle (2023).

It may still be surprising that people are less outspoken online than amongst their close friends, given the widespread hope that platforms would enhance freedom of expression for those who may not be comfortable sharing their political views with their offline friends and relations (Tufekci 2017). However, in a further exploratory analysis (see "× Likemindedness" model in Table 2), I interacted the platform treatment with an indicator for whether the respondent perceived their online network to be more or less likeminded than their close friends (see Appendix F.1). I recover a large and significant positive interaction effect: the self-censorship effect predicted in H1 is significant among the 46% of posters who perceive their offline networks to be more likeminded than their online networks, and also (with lesser magnitude) among the 38% who perceive their online and offline networks as equally likeminded, but the self-censorship effect is null among the 16% posters who perceive their online networks to be more likeminded than their offline networks.

So, platforms arguably do offer a refuge for political expression to people whose real-world friends might be hostile to their views, but this is a relatively small group, and the evidence indicates only that they do not *censor* their views online relative to with their close friend (and if their close friends are hostile to their perspective, this is not saying very much).

Variance Tests (H3 & H4) Pre-registered analyses for Hypotheses 3 and 4 employed an F-test for difference in variances. This test was chosen because Hypotheses 3 and 4 concern relative polarization (of posters relative to lurkers, and of posters-online relative to posters-offline, respectively), and because lexical ideology is (by construction) normally-distributed.

I tested Hypothesis 3 by comparing the variance of lurkers' and posters' close-friend lexical ideology, and found the variance of the latter to

TABLE 2. Platform Treatment Effects: Outspokenness (H1)

	Pooled	Facebook	Twitter	× Likemindedness
Intercept	0.27 (0.14)*	0.18 (0.21)	0.36 (0.18)*	0.38 (0.18)*
Platform Treatment	−0.30 (0.06)***	−0.27 (0.09)**	−0.34 (0.09)***	−0.58 (0.16)***
Age (Decades)	−0.03 (0.02)	−0.03 (0.03)	−0.02 (0.03)	−0.04 (0.02)
5-Point Ideology	−0.03 (0.04)	−0.02 (0.06)	−0.06 (0.06)	−0.02 (0.04)
7-Point Partisanship	0.03 (0.02)	−0.00 (0.03)	0.08 (0.03)*	0.02 (0.02)
College	−0.11 (0.06)	−0.02 (0.09)	−0.23 (0.09)*	−0.11 (0.07)
POC	0.36 (0.08)***	0.33 (0.11)**	0.38 (0.11)***	0.39 (0.08)***
Male	−0.01 (0.06)	0.12 (0.09)	−0.18 (0.09)	0.01 (0.07)
Likemindedness				−0.04 (0.04)
Platform Treatment × Likemindedness				0.12 (0.06)*
R^{2}	0.03	0.03	0.05	0.04
Adj. R^{2}	0.03	0.02	0.05	0.03
Num. obs.	1973	981	992	1834
RMSE	1.38	1.35	1.40	1.39

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

TABLE 3. Platform Treatment Effects: Lexical Ideology (H2)

	Pooled	Facebook	Twitter
Intercept	−1.99 (0.07)***	−1.83 (0.10)***	−2.12 (0.10)***
Platform Treatment	0.01 (0.03)	0.01 (0.04)	0.02 (0.04)
Age (Decades)	0.10 (0.01)***	0.09 (0.01)***	0.10 (0.01)***
5-Point Ideology	0.31 (0.02)***	0.28 (0.02)***	0.33 (0.03)***
7-Point Partisanship	0.15 (0.01)***	0.15 (0.01)***	0.15 (0.01)***
College	−0.18 (0.03)***	−0.24 (0.04)***	−0.13 (0.05)**
POC	0.04 (0.03)	−0.01 (0.05)	0.08 (0.05)
Male	0.16 (0.03)***	0.12 (0.04)**	0.20 (0.04)***
R^{2}	0.57	0.58	0.57
Adj. R^{2}	0.57	0.58	0.56
Num. obs.	1973	981	992
RMSE	0.63	0.58	0.67

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

FIGURE 4. Distribution of close-friend lexical ideology among lurkers *vs* posters (H3).

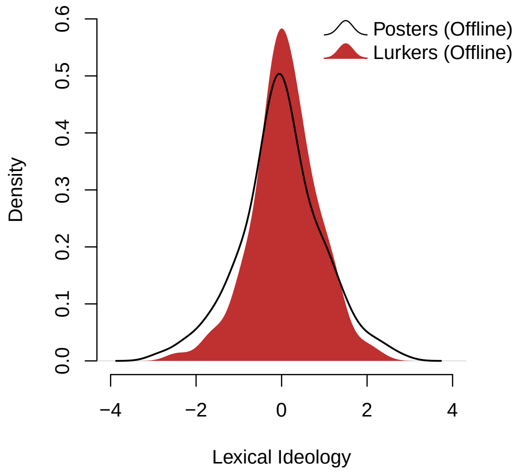
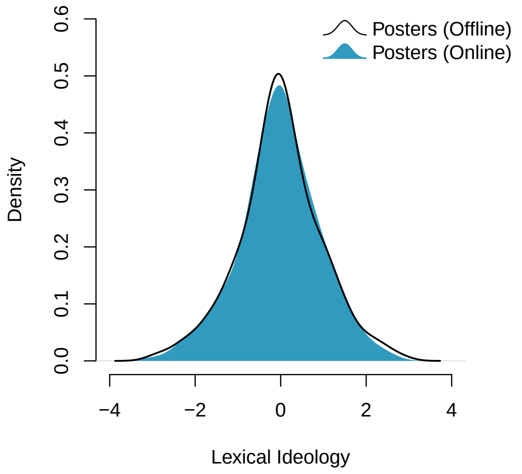


FIGURE 5. Distribution of posters' close-friend *vs* online lexical ideology (H4).



be significantly greater ($F = 1.58$, $p < 1 \times 10^{-6}$, pre-registered one-sided test). As seen in Figure 4, the distribution of posters' lexical ideology has greater density in the tails, and less density in the center, compared to lurkers. Exploratory analyses found that these results held for Facebook and Twitter users when analyzed separately ($F = 1.8$, $p < 1 \times 10^{-5}$, and $F = 1.41$, $p < 1 \times 10^{-2}$, respectively). If we use lurkers' offline speech as a proxy for what they self-censor online, this result is descriptively consistent with a self-censorship account of online polarization: people who post their political views on Twitter and Facebook have significantly more extreme offline speech patterns than users of these platforms who keep their political views to themselves on the internet.

I tested Hypothesis 4 by comparing the variance of posters' online lexical ideology to the variance of their close-friend lexical ideology, and found the difference to be null, as predicted ($F = 0.94$, $p = 0.85$, pre-registered one-sided test). As seen in Figure 5, the distributions of posters' online and offline lexical ideology hardly differ, which contradicts a preference-falsification account of online polarization. Exploratory analyses indicate that separately, Facebook and Twitter's polarization effects are both individually null ($F = 1.12$, $p \approx 0.08$, and $F = 0.8$, $p \approx 1$, respectively), at least at the planned threshold. That said, it is noteworthy that Facebook's polarization effect achieves what is conventionally considered marginal significance in the one-sided test that was pre-registered for H4, while Twitter's polarization effect arguably runs in the opposite direction: the variance of Twitter-posters' online lexical ideology is *narrower* than their offline lexical ideology, and a two-sided test finds this difference significant ($F = 0.8$, $p < 1 \times 10^{-2}$). So, while the pre-registered analyses give results consistent with the expectation of no platform polarization effect, exploratory analyses suggest a potential difference in this respect between the two platforms: Facebook may in fact polarize users' speech, and Twitter may actually *depolarize* users' speech.

DISCUSSION

This paper has applied an original method to explain the polarization of online political discourse. The evidence indicates this polarization is attributable to self-censorship on the part of moderate speakers, whose abstention from political speech distorts the distribution of opinions expressed online. I find little evidence of polarization by preference falsification.

Of course, there are other reasons why people might *perceive* online discourse as excessively polarized. For example, platforms' content ranking algorithms may promote extreme views, making them appear more prevalent than they really are. However, this possibility is beyond the scope of the present study, which is focused on individual-level mechanisms of online discourse polarization.

It is also possible that the polarization of online speech involves a dynamic process not fully captured in this one-shot experiment. For example, perhaps posters do engage in preference falsification when they first join a platform, but subsequently adjust their offline speech to match (perhaps due to cognitive dissonance). A longitudinal design would be needed to test this, and could also speak to the process by which moderate speakers select out of online political speech.

Overall, though, the present evidence clearly favors self-censorship over preference falsification. It is worth noting how starkly these findings contradict the overwhelming public consensus (Gallup 2022) that social platforms enable users to say things they wouldn't say offline – on the contrary, polarization appears to be driven by moderate users' *unwillingness* to post things that they *would* say offline. Polarization-by-self-censorship may be under-appreciated because it is somewhat counter-intuitive – like Holmes' dog-that-did-not-bark, the significance of silence only becomes apparent once we can infer *what is missing*.

This has important implications for the future conduct of research using social media data. Most importantly, such data should not be interpreted as a simple reflection of public opinion, but instead as "a heavily skewed tip of the iceberg" (Oswald et al. 2022). Scholars increasingly recognize that, in order to understand user-level behaviors on social media, it is necessary to collect data with a "user-centric" (Breuer et al. 2022) sampling frame that includes users who *do not do* the behavior that is the focus of the study.

It also has implications for how we understand the polarization of online discourse in particular. Social media users are not charlatans. Rather, the evidence I have gathered reveals a more sympathetic and even pitiable portrait of the typical user: a timid soul, cowed and alienated by a political discourse dominated by a cadre of brash ideologues, whose ire they fear to provoke, preferring instead to lurk in the shadows. I suspect that self-censorship dominates preference falsification in part because it is the path of least resistance for those who fear criticism: it's hard to talk about politics, and for most people, it is easier to stay silent than to falsify one's preferences. Though silence is a passive behavior, it nonetheless distorts the distribution of perspectives shared online, with significant consequences: if users' speech is unrepresentatively polarized, this could contribute to attitudinal and affective polarization of users themselves (e.g. Settle 2018).

Focusing on the mechanism of self-censorship also has practical implications for those who seek to depolarize online discourse. For one thing, it implies a need for a robust program of research to ascertain the reasons why certain people refrain from expressing themselves on social media platforms, to inform potential interventions.

For example, if users self-censor because they fear criticism, we should ask why people with moderate political perspectives might be especially fearful of criticism. One possibility is that moderate ideology is correlated with relevant psychological traits (again, see Bor and Petersen 2022). However, it is also plausible that moderates are structurally more vulnerable to criticism online: unlike strong ideologues, they may expect to be criticized by *both* left- and right-leaning users, effectively doubling the population of potential antagonists. Moderates may also be more disposed to *care* about criticism from both sides of the political spectrum, magnifying its psychological burden. One potential intervention to remedy this structural vulnerability would be a form of *enclave deliberation*[3] where moderates' political posts are shown preferentially to fellow moderates. Future research can test such interventions.

Another possibility is that the holders of moderate views feel less positive motivation to express them. For example, moderates may experience more cross-pressure (Lazarsfeld et al. 1948), and feel more conflicted about their political views – if this conflict connotes thoughtfulness, it is a shame that they do not contribute more to public discourse. On the other hand, perhaps these individuals simply *care less* about politics, in which case their abstention might actually be desirable. Future research can investigate both of these possibilities.

Framing the problems of online discourse in terms of *representation* may also offer a constructive new direction for public debate on this issue, which has for some time been stuck in an unhelpful dichotomy of *censorship versus freedom-of-speech*. While platforms have economic motives to take down content that reduces user engagement or advertising revenue (Klonick 2018), they do not necessarily have an incentive (or right) to delimit the range of legitimate ideological expression for their users, and any attempt to do so would likely attract accusations of politically-biased encroachments on user freedom.

If we conceptualize the problem in terms of representation, however, depolarization initiatives can actually *enhance* users' freedom of speech, by fostering more contributions from those who currently self-censor. Representation can also be defined concretely, relative to a thoughtfully-chosen reference point. For example, this paper takes close-friend conversation as its point of reference. This is not the only choice available, and one could certainly imagine a lively normative debate about what social media should be representative *of*, but this debate would be less obviously partisan (and so, hopefully, more productive) than attempting to police a boundary between "good" and "bad" political speech.

[3] Notably, this remedy is diametrically opposed to that implied by a preference falsification theory in which the presumed culprit is like-minded social pressure, as in Sunstein (2017), who recommends exposure to a *wider* range of perspectives as a remedy to polarization.

Ultimately, I hope the evidence I have presented helps to advance scholarship and public debate on the improvement of platformed discourse, by sharpening our understanding of polarization. We can, perhaps, rest easier in the knowledge that most users do not falsify the preferences they voice online. The task that stands before us is to understand why certain voices are missing.

REFERENCES

Bail, Christopher A. 2021. *Breaking the social media prism: how to make our platforms less polarizing*. Princeton: Princeton University Press.

Blair, Graeme, Jasper Cooper, Alexander Coppock, Macartan Humphreys, and Luke Sonnet. 2024. *estimatr: Fast Estimators for Design-Based Inference*. R package version 1.0.2, https://github.com/DeclareDesign/estimatr.

Bor, Alexander and Michael Bang Petersen. 2022. “The Psychology of Online Political Hostility: A Comprehensive, Cross-National Test of the Mismatch Hypothesis”. *American Political Science Review* 116 (1): 1–18.

Breuer, Johannes, Zoltán Kmetty, Mario Haim, and Sebastian Stier. 2022. “User-centric approaches for collecting Facebook data in the ‘post-API age’: experiences from two stu”. *Information, Communication & Society* 26 (14): 2649–2668.

Carlson, Taylor N. and Jaime E. Settle. 2023. “Freedom of Expression in Interpersonal Interactions”. *PS: Political Science & Politics* 56 (2): 245–249.

Conan Doyle, Arthur. 1892. “The Adventure of Silver Blaze”. *The Strand Magazine*.

Davison, W Phillips. 1983. “The Third-Person Effect in Communication”. *Public Opinion Quarterly* 47 (1): 1–15.

Fowler, Anthony, Seth J. Hill, Jeff Lewis, Chris Tausanovitch, Lynn Vavreck, and Christopher Warshaw. 2023. “Moderates”. *American Political Science Review* 117 (2): 643–660.

Gallup. 2022, March. “Media and Democracy: Unpacking America’s Complex Views on the Digital Public Square”. Technical report, Washington DC.

Goffman, Erving. 1956. *The Presentation of Self in Everyday Life*. Edinburgh: University of Edinburgh Press.

Haidt, Jonathan. 2022. “Why the Past 10 Years of American Life Have Been Uniquely Stupid”.

Hill, Jesse and James Fanciullo. 2023. “What’s wrong with virtue signaling?”. *Synthese* 201 (4): 117.

Kim, Jin Woo, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. “The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity”. *Journal of Communication* 71 (6).

Klonick, Kate. 2018. “The New Governors: The People, Rules, and Processes Governing Online Speech”. *Harvard Law Review* 131 (1598): 73.

Krippendorff, Klaus. 2004. *Content analysis: an introduction to its methodology* (2nd ed ed.). Thousand Oaks, Calif: Sage.

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

Lazarsfeld, Paul, Bernard Berelson, and Hazel Gaudet. 1948. *The People’s Choice: How the*

Voter Makes Up His Mind in a Presidential Campaign (2nd ed.). New York: Columbia University Press.

McGarty, Craig, John C. Turner, Michael A. Hogg, Barbara David, and Margaret S. Wetherell. 1992. “Group polarization as conformity to the prototypical group member”. *British Journal of Social Psychology* 31 : 1–20.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. “Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict”. *Political Analysis* 16 (4): 372–403.

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

Nonnecke, Blair and Jenny Preece. 2000. “Lurker demographics: counting the silent”. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, The Hague, pp. 73–80. ACM.

Oswald, Lisa, Simon Munzert, Pablo Barbera, Andrew Markus Guess, and JungHwan Yang. 2022. “Beyond the tip of the iceberg? Exploring Characteristics of the Online Public with Digital Trace Data”.

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

Slapin, Jonathan and Sven-Oliver Proksch. 2008. “A Scaling Model for Estimating Time-Series Party Positions from Texts”. *American Journal of Political Science* 52 (3).

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

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

Wojcik, Stefan and Adam Hughes. 2019, April. “Sizing Up Twitter Users”. Technical report, Pew Research Center.

APPENDIX A: EXPLORATORY RESEARCH: FALSIFICATION PERCEPTIONS

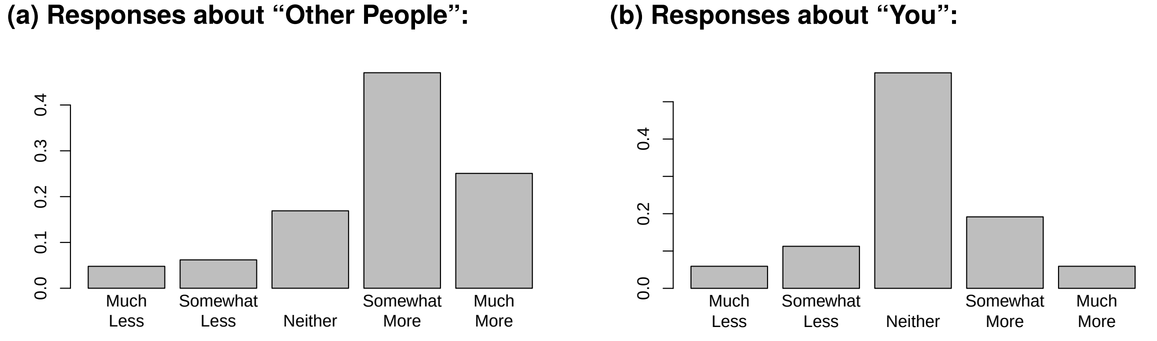
As part of the exploratory research that motivated this project, I conducted a survey of Democrat-identifying social media users (N=355, recruited from Amazon Mechanical Turk, Winter 2020) that posed direct questions about respondents perceptions of the prevalence of polarizing preference falsification on social media. In particular, I asked respondents two different versions of the same question:

1. “Do you think *other people* are more or less partisan on social media, compared to how *they* are in personal interactions?”
2. “Do you think *you* are more or less partisan on social media, compared to how *you* are in personal interactions?”

I presented both versions in randomized order, and the only difference was whether the question referred to the respondent themselves or to “other people.” As shown in Figure A1a, a large majority of respondents accused *other people* of exaggerating their partisanship online (72% of respondents said that others were “somewhat” or “much” more partisan online). But, as shown in Figure A1b, most respondents claimed that *they themselves* were no more or less partisan online than they were offline. This evidence appears consistent with the “third person effect” (Davison 1983) that has long been observed by scholars of political communication: individuals tend to believe that other people are more easily influenced (in the canonical case, by political messaging) than they themselves are. In this case, it appears that the average social media user believes that “other people” express a more partisan version of themselves on social media than they do in personal interactions, while maintaining that their own online speech is unaffected by any such distortion.

Of course, this analysis relies on an unrepresentative convenience sample, and so must be treated with caution. Nonetheless, this exploratory finding helps to motivate the study presented in the main text, in two ways: first, it indicates that preference falsification has popular credence as a theory of

FIGURE A1. Perceptions of Polarizing Preference Falsification


(a) Responses about “Other People”:

(b) Responses about “You”:

*“Do you think **other people** are more or less partisan on social media, compared to how **they** are in personal interactions?”*

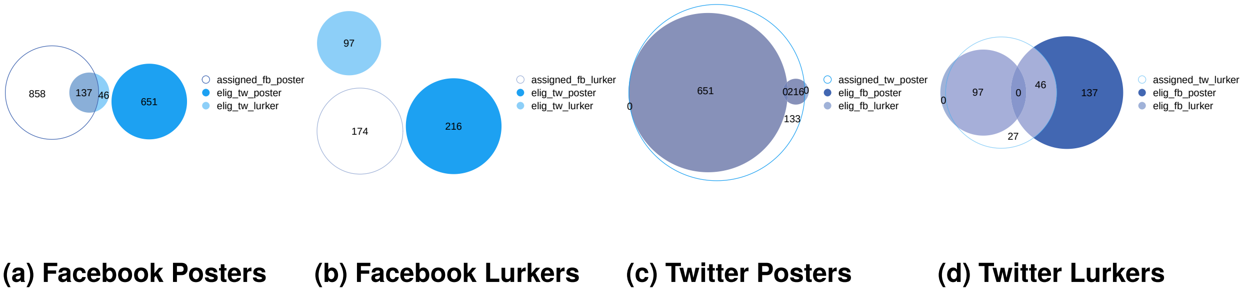
*“Do you think **you** are more or less partisan on social media, compared to how **you** are in personal interactions?”*

online polarization, in which nearly three-quarters of respondents report some level of belief – at least when applied to "other people." This makes the theory all the more worthy of empirical investigation, especially since if a large number of people subscribe to an incorrect theory of online polarization, they will struggle to identify or advocate for effective solutions to this major contemporary problem. Second, the fact that a majority of respondents *deny engaging in this behavior themselves* casts suspicion on this theory: the perception that individual speakers "over-represent their partisanship" online may simply be an illusion, derived from the over-representation of partisan speakers *as a group* in online discourse. This interpretation is consistent with the results of the study presented in the main text.

APPENDIX B: ELIGIBILITY

Participants who qualified for multiple groups were assigned to the least-filled group at time of recruitment. This was done to ensure sufficient sample size in each group, although it also means that multiply-eligible participants were preferentially allocated to the hardest-to-fill groups, namely Twitter Posters and Twitter Lurkers (see Figure B2). Importantly, this allocation procedure had no bearing on the subsequent treatment randomization, and so does not distort the pre-registered hypothesis tests. However, it does mean that exploratory analyses that compare the Twitter and Facebook samples against each other should be interpreted with some caution, since the Twitter sample over-represents dual-users.

FIGURE B2. Allocation of multiply-eligible participants (filled circles) to groups (outlined circle in each panel). Panel (a) thus shows that 137 participants who were categorized as Facebook Posters would also have been eligible as Twitter Lurkers. Panel (b) shows that no multiply-eligible participants were used as Facebook lurkers. Panels (c) and (d) shows that a large number of Twitter posters and lurkers, respectively, would also have been eligible as Facebook Posters or Lurkers.



(a) Facebook Posters (b) Facebook Lurkers (c) Twitter Posters (d) Twitter Lurkers

APPENDIX C: PILOT & POWER ANALYSIS

MTurk Pilot

I ran a pilot (N=798) of this experiment on MTurk in September 2021. Due to sample limitations, these analyses pooled users of Facebook, Twitter, and other platforms. Figure C3 shows the linear regression estimates of platforms' effects on both of these dimensions of speech (right panel). My pilot results indicated that platforms shift speech significantly *rightward and downward*. My pilot also found that platforms cause users to self-censor a wide variety of political phrases that they *would* say offline. I also find that this effect was moderated by like-mindedness: people whose online networks are more like-minded (than their close friends) were more outspoken online than off, but these are a minority. For most users, online platforms seemed to induce avoidance of political language.

To assess platform effects on the polarization of speech, I compared the variance of lexical ideology between the platform treatment and close-friend control (see Figure C4), and found no evidence that platforms polarize speech. Rather, I found evidence consistent with polarization by self-censorship: the users who speak up about politics online (the "posters") have significantly more extreme baseline (that is, close-friend) speech ideologies, compared to users who avoid talking about politics online (the "lurkers").

FIGURE C3. Coefficient plot platform effects on speech plotted in both dimensions, relative to the control condition of conversing with a close friend.

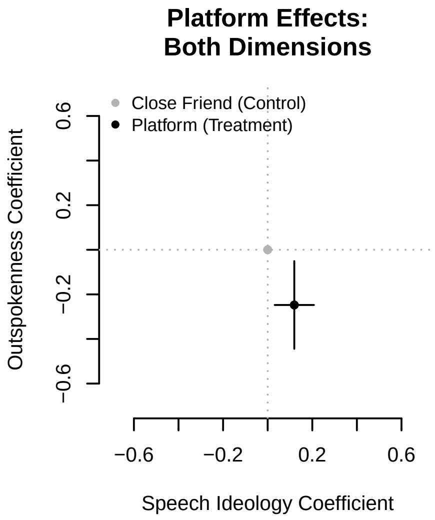
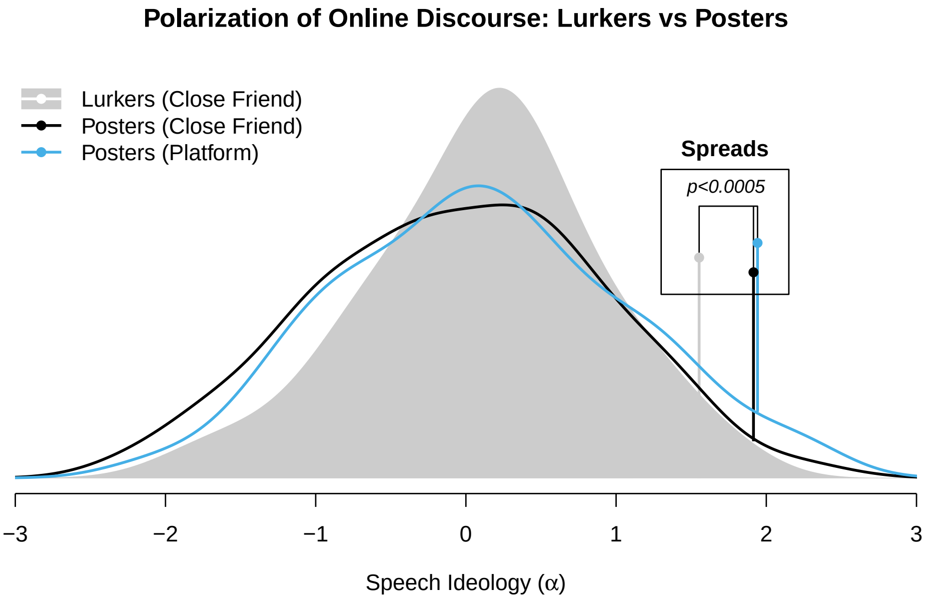


FIGURE C4. Density plots of respondent lexical ideology α, in three exhaustive subsets of the Study 2 data: lurkers speaking with a close friend (gray), posters speaking with a close friend (black), and posters posting on their preferred online platform (blue).



Power

To calculate minimum necessary sample sizes for my proposed analyses, I conducted power analyses by simulation, drawing synthetic datasets from my pilot data, and estimating the same models as I used in my pilot analyses to estimate treatment effects, but dropping variables, like issue ideology and ideological identity strength, which would not be included in a TESS survey. I then ran 10,000 simulations for each candidate sample size, and found that N = 994 would be sufficient to achieve 80% power to detect the left-right effect on lexical ideology that was observed in the pilot. Since this effect was smaller than the up-down effect observed on outspokenness, I assumed a minimum necessary N of **994 posters for each platform** to test both hypotheses.

I then estimated the necessary sample size of lurkers for testing for the descriptive difference-in-variances hypothesized in H3 by holding fixed the assumed poster N of 994, and simulating 10,000 tests for difference-in-variances at different synthetic sample sizes of lurkers. I found I would require a minimum sample of at least **96 lurkers for each platform** to achieve 80% power in this test. The power analysis conducted for Hypothesis 3 implies that the sample sizes calculated above would provide excellent power for detecting a polarization effect of magnitude comparable to that observed for Hypothesis 3 in the pilot, if such an effect were to exist for Hypothesis 4.

APPENDIX D: H1 REGRESSIONS SUBSET BY RELATIVE LIKEMINDEDNESS

TABLE D1. Platform Treatment Effect Heterogeneities: Outspokenness × Relative Likemindedeness

height	Offline > Online	Offline = Online	Offline < Online
Intercept	0.49 (0.22)*	0.13 (0.25)	0.05 (0.40)
Platform Treatment	−0.40 (0.09)***	−0.24 (0.11)*	0.04 (0.17)
Age (Decades)	−0.05 (0.03)	0.01 (0.04)	−0.07 (0.05)
5-Point Ideology	−0.03 (0.06)	0.01 (0.07)	−0.01 (0.10)
7-Point Partisanship	0.02 (0.04)	−0.00 (0.04)	0.08 (0.05)
College	−0.18 (0.09)	−0.16 (0.11)	0.21 (0.16)
POC	0.44 (0.12)***	0.32 (0.13)*	0.40 (0.17)*
Male	0.02 (0.09)	−0.03 (0.11)	0.03 (0.17)
R^2	0.06	0.02	0.04
Adj. R^2	0.05	0.01	0.02
Num. obs.	844	689	301
RMSE	1.33	1.43	1.44

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

APPENDIX E: RESEARCH ETHICS

This research adhered stringently to the ethical standards, and specifically to the Principles and Guidance for Human Subjects Research as set forth by the American Political Science Association. This research applied a methodology (the “What Would You Say?” question) designed to characterize participants’ speech with full consent, and without any risk of encroachment on participants’ privacy or confidentiality, and without employing any deception. Participants affirmatively volunteered all information they provided about their political speech, and no personally-identifiable information was collected. Overall, this methodology provides an exceptionally low-risk way of studying political speech. This research was approved by Princeton University IRB #15208. This research was supported by the TESS Special Competition Using Targeted Samples, with supplementary funding provided by the Center for the Study of Democratic Politics at Princeton University. The survey was fielded by NORC at the University of Chicago. Participants were compensated by NORC in the form of “AmeriPoints,” and amounts were determined by NORC, based on the amount of time the participant spent taking this and potentially other surveys. The researcher declares no conflicts of interest.

APPENDIX F: VARIABLES & QUESTIONNAIRE

This appendix contains the questionnaire implemented by NORC at the University of Chicago (project number 9089.121), per my instructions.

Likemindedness Question

Although full questionnaire details can be found below, I here highlight the question used to measure relative likemindedness of online and offline networks, for quick reference:

In general, are your political views more similar to your closest personal friends, or to the people you engage with on [Facebook/Twitter]?

○ Much more similar to my closest personal friends
○ Somewhat more similar to my closest personal friends
○ Equally similar to both
○ Somewhat more similar to the people I engage with on [Facebook/Twitter]
○ Much more similar to the people I engage with on [Facebook/Twitter]

Pre-Loaded Variables

The following tables describe variables that were pre-loaded into the survey (if available).

TABLE F2. Standard demographic preloaded variables.

Variable Name	Variable Type	Variable Label
S_AGE	Numeric	Age
S_GENDER	String	Gender
S_RACETH	Numeric	Race/ethnicity
S_EDUC	Numeric	Education
S_EDUC5	Numeric	5-level education
S_MARITAL	Numeric	Marital Status
S_EMPLOY	Numeric	Current employment status
S_INCOME	Numeric	Household income
S_HHINC_4	Numeric	4-level income
S_HHINC_9	Numeric	9-level income
S_STATE	String	State
S_METRO	Numeric	Metropolitan area flag
S_INTERNET	Numeric	Household internet access
S_HOUSING	Numeric	Home ownership
S_HOME_TYPE	Numeric	Building type of panelist's residence
S_PHONESERVC	Numeric	Telephone service for the household
S_HHSIZE	Numeric	Household size (including children)
S_HH01	Numeric	Number of HH members age 0-1
S_HH25	Numeric	Number of HH members age 2-5
S_HH612	Numeric	Number of HH members age 6-12
S_HH1317	Numeric	Number of HH members age 13-17
S_HH18OV	Numeric	Number of HH members age 18+
S_file_date	Date	Date
S_GENFRACE	Numeric	GenF custom race

TABLE F3. Standard sample preloaded variables.

Variable Name	Variable Type	Variable Label
Username	Numeric	Analogous to Member_PIN
P_Batch	Numeric	Batch Number (if only one assignment, then everyone will be 1)
Dialmode	Numeric	CATI Dialmode (predictive, preview, etc)
P_LCS	Numeric	Life cycle stage, 0=released but not touched
Y_FCELLP	String	
Surveylength	Numeric	Estimated length of survey
Incentwcomma	String	Study specific
P_Hold01	Numeric	Prevents dialing cases without phone numbers
PANEL_TYPE	Numeric	(1) AmeriSpeak (2) Next Generation (3) GenF Extended (not in use) (4) AmeriSpeak Teen Panel (11) UTUS Converted (20) Lucid (21) SSI (50) Household 13-17 (51) Household < 13 (52) Household Adult

TABLE F4. Custom survey-specific preloaded variables.

Variable Name	Variable Type	Variable Label
S_PARTY7ID	Numeric	(1) Strong Democrat, (2) Moderate Democrat, (3) Lean Democrat, (4) Don't Lean/independent/None, (5) Lean Republican, (6) Moderate Republican, (7) Strong Republican
S_IDEO20	Numeric	(1) Very liberal, (2) Somewhat liberal, (3) Moderate, (4) Somewhat conservative, (5) Very conservative,
P_RELIG	Numeric	(1) Protestant (Baptist, Methodist, Non-denominational, Lutheran, Presbyterian, Pentecostal, Episcopalian, Reformed, Church of Christ, Jehovah's Witness, etc.), (2) Roman Catholic (Catholic), (3) Mormon (Church of Jesus Christ of Latter-day Saints/LDS), (4) Orthodox (Greek, Russian, or some other orthodox church), (5) Jewish (Judaism), (6) Muslim (Islam), (7) Buddhist, (8) Hindu, (9) Atheist (do not believe in God), (10) Agnostic (not sure if there is a God), (11) Nothing in particular, (12) Just Christian, (13) Unitarian (Universalist), (14) Other, please specify
P_RELIG_OE	STRING	[TEXTBOX]
P_ATTEND	Numeric	(1) Never, (2) Less than once per year, (3) About once or twice a year, (4) Several times a year, (5) About once a month, (6) 2-3 times a month, (7) Nearly every week, (8) Every week, (9) Several times a week
P_PHRASE_7 to P_PHRASE_20	Numeric	CREATE 14 VARIABLES P_PHRASE_7 to P_PHRASE_20 *NO VALUES 1-6* (7) Black lives matter, (8) abortion is healthcare, (9) cis-gender, (10) privilege, (11) climate crisis, (12) toxic masculinity, (13) defund the police, (14) equity, (15) empathy, (16) micro-aggression, (17) safe space, (18) POC, (19) words matter, (20) eat the rich, (21) mansplain, (22) heteronormative, (23) voter suppression, (24) all lives matter, (25) sanctity of life, (26) reverse racism, (27) libtard, (28) patriot, (29) illegal alien, (30) traditional values, (31) blue lives matter, (32) mainstream media, (33) thug, (34) do your own research, (35) MAGA, (36) free speech, (37) cancel culture, (38) personal responsibility, (39) biological women, (40) voter fraud

Questionnaire

The full questionnaire and coding instructions are printed below. Note that this includes demographic questions to be asked if the relevant variables could not be pre-loaded.

[SHOW ALL] [DISPLAY – WINTRO_1]
[CAWI] Thank you for agreeing to participate in our new AmeriSpeak survey!
[ALL] This survey is about politics, and asks questions about whether and how you talk about politics.
[CAWI] To thank you for sharing your opinions, we will give you a reward of [INCENTWCOMMA] AmeriPoints after completing the survey. As always, your answers are confidential.
[CAWI] Please use the "Continue" button to move forward within the questionnaire. Do not use your browser buttons.

[SHOW IF PANEL_TYPE>=20]
DISPLAY – OPTINTRO.
Thank you for agreeing to participate in our survey! This survey is about politics and asks questions about whether and how you talk about politics. Your answers are confidential.
Please use the "Continue" button to navigate between the questions within the questionnaire. Do not use your browser buttons.

[SHOW IF PANEL_TYPE>=20]
[NUMBOX]
[FORCE RESPONSE: “Please enter in your age. We require this information for your responses to be counted”]
AGE2.
What is your current age?
[0-100] years
[IF AGE2<18 OR AGE2>24, TERMINATE AND SET QUAL=2]
[COMPUTE S_AGE=AGE2]

[SHOW IF PANEL_TYPE>=20]
[SP]
[FORCE RESPONSE: “Please tell us your gender. We require this information for your responses to be counted”]
GENDER2.
Are you . . . .
RESPONSE OPTIONS:

- Male
- Female

[COMPUTE S_GENDER=GENDER2]

[SHOW IF PANEL_TYPE>=20]

[FORCE RESPONSE]
[SP]
HHSIZE1.
Tell us a little about your household. <u>Including yourself</u>, how many persons currently live in your household at least 50 percent of the time? Please include any children as well as adults.
RESPONSE OPTIONS:

- One person, I live by myself
- Two persons
- Three persons
- Four persons
- Five persons
- Six or more persons

[COMPUTE S_HHSIZE=HHSIZE1]

[SHOW IF HHSIZE1>1]
[FORCE RESPONSE]
[NUMBOXES]
Please tell us how many persons currently living in your household, including yourself, are. . .
HH01S. ___ 0-1 years old
HH25S. ___ 2-5 years old
HH612S. ___ 6-12 years old
HH1317S. ___ 13-17 years old
HH18OVS. ___ 18 years old or older
HHtotal. ____ Total household members
HHtotal SHOULD SHOW AUTO-SUM OF HH01S-H18OVS
DO NOT ALLOW R TO CONTINUE IN SURVEY IF HHtotal<HHSIZE1
COMPUTE HH01=HH01S .
COMPUTE HH25=HH25S .
COMPUTE HH612=HH612S .
COMPUTE HH1317=HH1317S .
COMPUTE HH18OV=HH18OVS .
COMPUTE HHMINORS=sum(HH01, HH25, HH612, HH1317)

[SHOW IF PANEL_TYPE>=20]
[NUMBOX] [FORCE RESPONSE]
ZIP.
What is your zipcode?
__[00000-99999,777777,999998,999999]__
[ZIP validation check: must contain 5-digits, only numbers, leading 0s okay]

[SHOW IF PANEL_TYPE>=20]
[DROPDOWN] [FORCE RESPONSE]
STATE2.
What state do you live in?
[DROPDOWN LIST OF STATES]

[COMPUTE S_STATE=STATE2]

[SHOW IF PANEL_TYPE>=20]
[SP] [FORCE RESPONSE]
[custom prompt: "Information about any possible Hispanic ethnicity is very important. We greatly appreciate your response to this question."]
HISPAN.
This question is about Hispanic ethnicity. Are you of Spanish, Hispanic, or Latino descent?
RESPONSE OPTIONS:

- No, I am not
- Yes, Mexican, Mexican-American, Chicano
- Yes, Puerto Rican
- Yes, Cuban
- Yes, Central American
- Yes, South American
- Yes, Caribbean
- Yes, Other Spanish/Hispanic/Latino

[SHOW IF PANEL_TYPE>=20]
[MP] [FORCE RESPONSE]
RACE_1.
Please indicate what you consider your racial background to be. We greatly appreciate your help. The categories we use may not fully describe you, but they do match those used by the Census Bureau.
Please check one or more categories below to indicate what <u>race or races</u> you consider yourself to be.
RESPONSE OPTIONS:

1. White
2. Black or African American
3. American Indian or Alaska Native – <i>Type in name of enrolled or principal tribe.</i> [TEXTBOX]
4. Asian Indian
5. Chinese
6. Filipino
7. Japanese
8. Korean
9. Vietnamese
10. Other Asian – <i>Type in race</i> [TEXTBOX]
11. Native Hawaiian
12. Guamanian or Chamorro
13. Samoan
14. Other Pacific Islander – <i>Type in race</i> [TEXTBOX]
15. Some other race – <i>Type in race</i> [TEXTBOX]

[SHOW IF PANEL_TYPE>=20]
DISPLAY - HHINCINTRO.

The next question is about the <u>total income</u> of YOUR HOUSEHOLD for [CURRENTYEAR-1]. Please include your own income PLUS the income of all members living in your household (including cohabiting partners and armed forces members living at home). Please count income BEFORE TAXES and from all sources (such as wages, salaries, tips, net income from a business, interest, dividends, child support, alimony, and Social Security, public assistance, pensions, or retirement benefits).

[SHOW IF PANEL_TYPE>=20]
[SP]
[FORCE RESPONSE] Information about your household income is very important. We greatly appreciate your response and will keep your answer confidential.]
INCOME2.
Was your total HOUSEHOLD income in [CURRENTYEAR-1]. . .
RESPONSE OPTIONS:

- Less than $5,000
- $5,000 to $9,999
- $10,000 to $14,999
- $15,000 to $19,999
- $20,000 to $24,999
- $25,000 to $29,999
- $30,000 to $34,999
- $35,000 to $39,999
- $40,000 to $49,999
- $50,000 to $59,999
- $60,000 to $74,999
- $75,000 to $84,999
- $85,000 to $99,999
- $100,000 to $124,999
- $125,000 to $149,999
- $150,000 to $174,999
- $175,000 to $199,999
- $200,000 or more

[COMPUTE S_INCOME=INCOME2]
IF INCOME2=1-6 S_HHINC4=1
IF INCOME2=7-10 S_HHINC4=2
IF INCOME2=11-13 S_HHINC4=3
IF INCOME2=14-18 S_HHINC4=4
IF INCOME2=1-2 S_HHINC9=1
IF INCOME2=3-4 S_HHINC9=2
IF INCOME2=5-6 S_HHINC9=3
IF INCOME2=7-8 S_HHINC9=4
IF INCOME2=9 S_HHINC9=5
IF INCOME2=10-11 S_HHINC9=6
IF INCOME2=12-13 S_HHINC9=7
IF INCOME2=14-15 S_HHINC9=8
IF INCOME2=16-18 S_HHINC9=9

[SHOW IF PANEL_TYPE>=20]
[SP] [FORCE RESPONSE]
HOME_TYPE2.
Which best describes the building where you live?
RESPONSE OPTIONS:

- A one-family house detached from any other house
- A one-family house attached to one or more houses
- A building with 2 or more apartments
- A mobile home or trailer
- Boat, RV, van, etc

[COMPUTE S_HOME_TYPE=HOME_TYPE2]

[SHOW IF PANEL_TYPE>=20]
[SP] [FORCE RESPONSE]
HOUSING2.
Share with us a little about where you live. Are your living quarters. . .
RESPONSE OPTIONS:

- Owned or being bought by you or someone in your household
- Rented for cash
- Occupied without payment of cash rent

[COMPUTE S_HOUSING=HOUSING2]

[SHOW IF PANEL_TYPE>=20]
[SP] [FORCE RESPONSE]
Q5PHONE.
What best describes your telephone service for your household?
RESPONSE OPTIONS:

- Landline telephone only
- Have a landline, but mostly use cellphone
- Have cellphone, but mostly use landline
- Cellphone only

[COMPUTE S_PHONESERV=Q5PHONE]

[SHOW IF PANEL_TYPE=>20]
[SP] [FORCE RESPONSE]
ATTENTION.
Below is a list of numbers. Please select the number seven.
RESPONSE OPTIONS:

- 1
- 3
- 5
- 7
- 9

- 11
- 12

[IF ATTENTION<>4, TERMINATE AND SET QUAL=2]

[SHOW IF PANEL_TYPE>=20]
[SP] [FORCE RESPONSE]
MARITAL2.
Are you . . . .
RESPONSE OPTIONS:

- Married
- Widowed
- Divorced
- Separated
- Never married

[COMPUTE S_MARITAL=MARITAL2]

[SHOW IF PANEL_TYPE>=20]
[SP] [FORCE RESPONSE]
EDUC2.
What is the highest level of school you have completed?
RESPONSE OPTIONS:

1. No formal education
2. 1st, 2nd, 3rd, or 4th grade
3. 5th or 6th grade
4. 7th or 8th grade
5. 9th grade
6. 10th grade
7. 11th grade
8. 12th grade no diploma
9. High school graduate – high school diploma or the equivalent (GED)
10. Some college, no degree
11. Associate degree
12. Bachelor’s degree
13. Master’s degree
14. Professional or Doctorate degree

[COMPUTE S_EDUC=EDUC2]
IF EDUC2=1-8 COMPUTE S_EDUC5=1
IF EDUC2=9 COMPUTE S_EDUC5=2
IF EDUC2=10-11 COMPUTE S_EDUC5=3
IF EDUC2=12 COMPUTE S_EDUC5=4
IF EDUC2=13-14 COMPUTE S_EDUC5=5

[SHOW IF PANEL_TYPE>=20]
[SP] [FORCE RESPONSE]
EMPLOY2.

Which statement best describes your current employment status?
RESPONSE OPTIONS:

1. Working – as a paid employee
2. Working – self-employed
3. Not working – on temporary layoff from a job
4. Not working – looking for work
5. Not working – retired
6. Not working – disabled
7. Not working – other

[COMPUTE S_EMPLOY=EMPLOY2]

[SHOW IF PANEL_TYPE>=20]
[NUMBOX] [FORCE RESPONSE]
AGECONFIRM.
What year were you born?
[NUMBOX: 0-2022]
PN: TERMINATE AND SEND TO TERMSORRY IF (2022 – AGECONFIRM) > (AGE2 + 2) OR (2022 – AGECONFIRM) < (AGE2 - 2)

[SHOW IF PANEL_TYPE>=20]
TERMSORRY_OFF.
Thank you for your time today. Unfortunately you are not eligible for this study. We appreciate your participation.
SET QUAL=2 AND REDIRECT TO OPT-IN VENDOR
]
PM: PLEASE MAKE SURE THE DATE TIME RULE ALWAYS FOLLOWS FIRST QUESTION
INSERT ITEM TIMESTAMPS: TIME_FIRST, DATE_FIRST

MAIN SURVEY STARTS HERE

#[GRID; SP]
[DOUBLE PROMPT]
Q1.
Do you use any of the following social media platforms?
GRID ITEMS, RANDOMIZE:

- Facebook
- Snapchat
- TikTok
- Instagram
- WhatsApp
- Discord
- Twitter
- YouTube
- BeReal
- Mastodon

RESPONSE OPTIONS:

- Yes
- No

[SHOW IF ANY(Q1A-Q1J=1)]
[DOUBLE PROMPT]
[SPECIAL GRID; SP]
Q2.
Do you use any of the following social media platforms to post your opinions about politics or current events?
[CAWI] Please select all that apply.
[CATI] SELECT ALL THAT APPLY.
GRID ITEMS:

- [SHOW IF Q1A=1] Facebook
- [SHOW IF Q1B=1] Snapchat
- [SHOW IF Q1C=1] TikTok
- [SHOW IF Q1D=1] Instagram
- [SHOW IF Q1E=1] WhatsApp
- [SHOW IF Q1F=1] Discord
- [SHOW IF Q1G=1] Twitter
- [SHOW IF Q1H=1] YouTube
- [SHOW IF Q1I=1] BeReal
- [SHOW IF Q1J=1] Mastodon

RESPONSE OPTIONS:

- Yes
- No

PROGRAMMING: CREATE DATA-ONLY VARIABLE: QUOTA_DOV_ELIG [MP]
1=Facebook poster
2=Facebook lurker
3=Twitter poster
4=Twitter lurker
9=not eligible
IF Q2A=1 QUOTA_DOV_ELIG=1 ‘Facebook poster’
IF Q2A=2 QUOTA_DOV_ELIG=2 ‘Facebook lurker’
IF Q2G=1 QUOTA_DOV_ELIG=3 ‘Twitter poster’
IF Q2G=2 QUOTA_DOV_ELIG=4 ‘Twitter lurker’
ELSE QUOTA_DOV_ELIG=9
DISPLAY QUOTA_DOV_ELIG ON TESTING ONLY PAGE FOR CHECK PURPOSES
PROGRAMMING NOTE: USE QUOTA FUNCTIONALITY IN A4S AND VOXCO, NOT SYNCHED, ACTIVATE QUOTAS IN VCC AND A4S.
CREATE DATA-ONLY VARIABLE: DOV_ASSIGNED [SP]
1=Facebook poster
2=Facebook lurker
3=Twitter poster
4=Twitter lurker

9=Not assigned
*quota targets are a little higher to account for cleaning
CHECK IF DOV_ELIGIBLE QUOTA GROUP IS OPEN,
IF YES DOV_ASSIGNED=DOV_ELIGIBLE
IF MORE THAN ONE DOV_ELIGIBLE, ASSIGN TO LEAST FILLED ELIGIBLE OPEN BUCKET
IF QUOTA BUCKET FULL, SET TO OUT OF QUOTA AND SET DOV_ASSIGNED=9
IF DOV_ELIG=9 OR DOV_ASSIGNED=9, TERMINATE AND GO TO QUOTA_MET

[SHOW IF ASSIGNED=9] [REMOVE PREVIOUS BUTTON]
[DISPLAY - QUOTA_MET]
Thank you for your interest in our survey. At this time we have reached the desired number of completed interviews. Thank you and have a great day!

[SHOW IF PANEL_TYPE<20] We will redirect you to the AmeriSpeak Member Portal in n seconds.
EXIT AS QUOTA MET/CLOSED
PANEL_TYPE<20 auto-redirect to MEMBER PORTAL in 10 seconds, display remaining number of seconds in [n]
IF PANEL_TYPE>=20 REDIRECT TO
Measuring Relative Like-Mindedness of Close Friends vs Online Networks

---

#[SHOW IF QUOTA_DOV_ELIG=1 or 2]
[SP]
Q4.
In general, are your political views more similar to your <u>closest personal friends</u>, or more similar to the people you engage with on <u><i>Facebook<i/></u>?
RESPONSE OPTIONS:
SHOW IF RND_01=0; 1-5
SHOW IF RND_01=1; 5-1

- Much more similar to my closest personal friends
- Somewhat more similar to my closest personal friends
- Equally similar to both
- Somewhat more similar to the people I engage with on Facebook
- Much more similar to the people I engage with on Facebook

#[SHOW IF QUOTA_DOV_ELIG=3 or 4]
[SP]
Q5.
In general, are your political views more similar to your <u>closest personal friends</u>, or more similar to the people you engage with on <u><i>Twitter<i/></u>?
RESPONSE OPTIONS:
SHOW IF RND_01=0; 1-5
SHOW IF RND_01=1; 5-1

- Much more similar to my closest personal friends
- Somewhat more similar to my closest personal friends

- Equally similar to both
- Somewhat more similar to the people I engage with on Twitter
- Much more similar to the people I engage with on Twitter

The “What Would You Say?” Question
Shown to all respondents (but with different versions depending on Twitter/Facebook Lurker/Poster, and amongst posters there is a randomized treatment)
NOTE: It is important to record which phrases were sampled and displayed to each respondent, and to record the order in which the phrases were displayed.

[RECORD TIME SPENT ON SCREEN]

CREATE DATA ONLY VARIABLE FOR DOV_CONTEXT
1= with a close friend, who knows you very well
2= on Facebook
3= on Twitter
IF DOV_ASSIGNED=2 OR 4 DOV_CONTEXT=1
IF DOV_ASSIGNED=1 AND RND_00=0 DOV_CONTEXT=1
IF DOV_ASSIGNED=1 AND RND_00=1 DOV_CONTEXT=2
IF DOV_ASSIGNED=3 AND RND_00=0 DOV_CONTEXT=1
IF DOV_ASSIGNED=3 AND RND_00=1 DOV_CONTEXT=3

PROGRAMMING NOTE: Please make sure phrases are presented within double quotes separated by spaces.

#[SHOW ALL]
[GRID; 6,5,4,5: SP]
Q6.
<u>Here is a list of words and phrases</u> that someone might use when talking about politics.
[SPACE]
Please indicate whether each word/phrase is something <u>you would use [DOV_CONTEXT] </u>.
[SPACE]
<UNBOLD>Note: Please only consider whether you would use a phrase <u>sincerely</u>. 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.</UNBOLD>
GRID ITEMS, ALWAYS SHOW A-F ON FIRST SCREEN AND RANDOMIZE ITEMS WITHIN A-F; RANDOMIZE ITEMS G-T AND RECORD ORDER ACROSS SCREENS 2,3,4:

- “ systemic racism ”
- “ big government ”
- “ human rights ”
- “ America first ”
- “ LatinX ”
- “ snowflake ”
- “ [SHOW P_PHRASE_7] ”

- “ [SHOW P_PHRASE_8] ”
- “ [SHOW P_PHRASE_9] ”
- “ [SHOW P_PHRASE_10] ”
- “ [SHOW P_PHRASE_11] ”
- “ [SHOW P_PHRASE_12] ”
- “ [SHOW P_PHRASE_13] ”
- “ [SHOW P_PHRASE_14] ”
- “ [SHOW P_PHRASE_15] ”
- “ [SHOW P_PHRASE_16] ”
- “ [SHOW P_PHRASE_17] ”
- “ [SHOW P_PHRASE_18] ”
- “ [SHOW P_PHRASE_19] ”
- “ [SHOW P_PHRASE_20] ”

RESPONSE OPTIONS:

- Definitely <u>Would</u> Say
- Probably <u>Would</u> Say
- Probably <u>Wouldn’t</u> Say
- Definitely <u>Wouldn’t</u> Say

CREATE DATA ONLY VARIABLE FOR DOV_PLATFORM [MP]
IF MORE THAN ONE DOV_PLATFORM, ASSIGN TO 50-50 PROBABILITY BASED ON RND_02
1= Facebook
2= Twitter
IF Q1A=1 DOV_PLATFORM=1 ‘Facebook’
IF Q1G=1 DOV_PLATFORM=2 ‘Twitter’
IF Q1A=01 AND Q1G=01 AND RND_02=0 DOV_PLATFORM=’Facebook’
IF Q1A=01 AND Q1G=01 AND RND_02=1 DOV_PLATFORM= ‘Twitter’

#[SHOW ALL]
[MP]
Q7.
My account on [DOV_PLATFORM]...
[CAWI - REMOVE BOLD] <i> Please select all that apply. </i>
[CATI] SELECT ALL THAT APPLY.

- Shows my real name
- Shows a photo of my face
- Is visible to my real-world friends
- Is visible to my family
- Is visible to my boss, co-workers, or colleagues

#[SHOW IF (MISSING (S_IDEO20)) OR (MISSING (P_RELIG OR (IF P_RELIG=14 AND P_RELIG_OE IS MISSING)) OR (MISSING (P_ATTEND))]
[DISPLAY]
DEMO_INTRO.

Before we wrap up, just some quick background questions.

#[SHOW IF MISSING (S_PARTY7ID)]
[SP]
PID1.
Do you consider yourself a Democrat, a Republican, an Independent or none of these?
RESPONSE OPTIONS:

- Democrat
- Republican
- Independent
- None of these

#[SHOW IF PID1=1]
[SP]
PIDA.
Do you consider yourself a strong or not so strong Democrat?
RESPONSE OPTIONS:

- Strong Democrat
- Not so strong Democrat

#[SHOW IF PID1=2]
[SP]
PIDB.
Do you consider yourself a strong or not so strong Republican?
RESPONSE OPTIONS:

- Strong Republican
- Not so strong Republican

#[SHOW IF PID1=3, 4, 77, 98, 99]
[SP]
PIDi.
Do you lean more toward the Democrats or the Republicans?
RESPONSE OPTIONS:

- Lean Democrat
- Lean Republican
- Don’t lean

#[SHOW IF MISSING (S_IDEO20)]
[SP]
D3.
Generally speaking, do you consider yourself to be a liberal, moderate, or conservative?
RESPONSE OPTIONS:

- Liberal

- Moderate
- Conservative

#[SHOW IF D3=1]
[SP]
D4.
Do you consider yourself:
RESPONSE OPTIONS:

- Very liberal
- Somewhat liberal

#[SHOW IF D3=3]
[SP]
D5.
Do you consider yourself:
RESPONSE OPTIONS:

- Very conservative
- Somewhat conservative

#[SHOW IF MISSING P_RELIG OR (IF P_RELIG=14 AND P_RELIG_OE IS MISSING)]
RELIG.
What is your present religion, if any?
RESPONSE OPTIONS:

- Protestant (Baptist, Methodist, Non-denominational, Lutheran, Presbyterian, Pentecostal, Episcopalian, Reformed, Church of Christ, Jehovah's Witness, etc.)
- Roman Catholic (Catholic)
- Mormon (Church of Jesus Christ of Latter-day Saints/LDS)
- Orthodox (Greek, Russian, or some other orthodox church)
- Jewish (Judaism)
- Muslim (Islam)
- Buddhist
- Hindu
- Atheist (do not believe in God)
- Agnostic (not sure if there is a God)
- Nothing in particular
- Just Christian
- Unitarian (Universalist)
- Something else, please specify: [TEXTBOX]

#[SHOW IF MISSING P_ATTEND]
ATTEND.
How often do you attend religious services?
RESPONSE OPTIONS:

- Never

- Less than once per year
- About once or twice a year
- Several times a year
- About once a month
- 2-3 times a month
- Nearly every week
- Every week
- Several times a week

RE-COMPUTE QUAL=1 “COMPLETE”
SET CO_DATE, CO_TIME, CO_TIMER VALUES HERE
CREATE MODE_END
1=CATI
2=CAWI
