Breaking the Spiral of Silence*

Yuen Ho†

Yihong Huang‡

March 21, 2026

Abstract

Self-censorship is pervasive in today's social climate, distorting political discourse and decision-making processes that either explicitly or implicitly take public opinion into account. What drives self-censorship, and how can misperceptions about public opinion be corrected? The Spiral of Silence theory offers insights, positing that people who hold views perceived as socially inappropriate tend to self-censor, generating a distribution of expressed views that is skewed towards appropriate opinions. If the attention paid to silence is limited, this can exacerbate self-censorship and create an equilibrium where only socially appropriate views are expressed and considered dominant. Conceptually, we develop a two-stage signaling model in which self-censorship and limited attention to silence jointly interact to affect beliefs and public expression. We then experimentally test our model predictions in a dynamic, unified experimental framework where college students discuss controversial topics together. Our results show that students with socially inappropriate views self-censor to a significant degree. Given the limited attention paid to silence in the status quo, self-censorship amplifies over time. We experimentally increase the salience of silence and find that this leads to both more accurate beliefs about public opinion and increases the diversity of views expressed. Because inference and expression are self-reinforcing, different levels of attention to silence lead to diverging equilibria.

*We thank Benjamin Enke, Matthew Rabin, Gautam Rao, and David Yang for help and advice throughout different stages of this project. We thank the editor, Leonardo Bursztyn, and four anonymous referees for detailed comments and helpful suggestions. We would also like to thank Stefano DellaVigna, Edward Glaeser, David Laibson, Jeremy Magruder, Juanjuan Meng, Ricardo Perez-Truglia, Vincent Pons, Jesse Shapiro, Tomasz Strzalecki, and participants at presentations and conferences at Harvard University, UC Berkeley, and Stanford University for valuable comments and discussions. We thank Sasha Michelle Ballesteros, Mahima Bansal, Madison Butolph, Cate Sinclair Carlson, Evelyn Cheng, and Shuai Yuan for excellent research assistance. This project was approved by the Institutional Review Board at Harvard University (IRB22-0757) and at University of California, Berkeley (2022-11-15819 and 2023-08-16614). The experiment was pre-registered in the AEA registry (AEARCTR-0009505). Financial support from the Kenneth C. Griffin Economics Research Funds and UC Berkeley Xlab is gratefully acknowledged. All errors are our own.

†Trinity College Dublin, Department of Economics. Email: ho.yuen@tcd.ie

‡Peking University, Guanghua School of Management. Corresponding Author, Email: yihongh@gsm.pku.edu.cn.

1 Introduction

Political discourse is often characterized by a vocal minority and a silent majority. In American history, President Richard Nixon coined the term "silent majority" to describe Americans not involved in the visible counterculture and anti-war movements of that era. President Donald Trump later invoked the same term to describe his supporters during his electoral campaigns.[1] Examples of the "silent majority" abound across various platforms: only a minority of town hall attendees raise comments, and on social media platforms like Twitter, 10% of users account for 80% of all tweets (Pew Research Center, 2019).

In today's political climate, characterized by polarized views and an alleged "cancel culture," self-censorship may be especially pronounced. A majority of Americans (62%) say that they have political opinions that they are afraid to share, and this is true for both Democrats and Republicans (Ekins, 2020). In particular, there has been recent media attention about self-censorship becoming more prevalent on college campuses, where strong social norms prevail. College students frequently sidestep controversial topics, avoid "triggering" terms, and cancel speakers with contentious viewpoints (New York Times, 2022). In fact, 80% of college students report that they self-censor at least some of the time (College Pulse, 2021).

Freedom of expression is often an intrinsic political value and is believed to be important for fostering learning and innovation (Coudray, 2015). From a policy perspective, self-censorship is also important because many decision making processes either explicitly or implicitly take public opinion into account.[2] In an environment where individuals self-censor due to potential social backlash, the views that are expressed will not accurately represent the views that the public actually holds, leading to biased collective decisions.

In a dynamic context, the effects of perceived social norms and political expression can be self-reinforcing, creating a "spiral of silence" (Noelle-Neumann, 1974). The spiral is initiated by social norms fostering the expression of a singular acceptable view, which results in self-censorship among those who hold differing views. Because silence is easily overlooked, people may treat the distribution of expressed views as representative. This, in turn, causes them to overestimate the popularity of the socially appropriate view, which propels further self-censorship. Ultimately, this can create a self-reinforcing equilibrium where only socially appropriate views are expressed and believed to be dominant.

In this paper, we ask two questions. First, does the spiral of silence exist in practice? That is, do people self-censor when they hold socially inappropriate views and does this amplify when attention to silence is limited? Second, how can we break the spiral of silence?

[1] See Nixon's "silent majority" speech (C-SPAN, 1969) and Trump's campaign report (WBUR, 2020).

[2] Classic theories of democratic representation propose that politicians aim to support their constituents' preferences over policies (for example, see (Przeworski, Stokes, and Manin, 1999))

Specifically, we study the effect of increasing attention to silence on beliefs about others, public expression, and the resulting equilibrium of perceived social norms. We focus on the particular role of silence in the creation and perpetuation of these dynamics, as silence may be especially overlooked in an attention-based world.

To formalize the spiral of silence and motivate our experiment, we develop a two-stage signaling model in which individuals hold either socially appropriate or inappropriate views. In the first stage, information senders make public expression decisions by weighing the intrinsic value of honest expression against potential social backlash and the psychological cost of lying. Crucially, the perceived social sanction costs of expressing socially inappropriate views depend on the perceived dominance of appropriate views within society, while expressing appropriate views incurs no such costs. Thus, individuals with inappropriate views, expecting sanctions, are more likely to self-censor. Information receivers observe the publicly expressed views and silence and use these signals to update their beliefs about others' views. For tractability, we assume that receivers update in a Bayesian way to the signals they attend to.[3] We further distinguish between two types of receivers with different levels of attention to silence. A fraction of receivers pay attention to silence and correctly infer that dissenters are more likely to stay silent, whereas inattentive receivers ignore silence and treat expressed views as representative. In the second stage, receivers become senders and make their own expression decisions based on their updated beliefs.

The model predicts that receivers who pay attention to silence form lower estimates of the prevalence of socially appropriate views and are correspondingly more willing to express opposing views as senders. Notably, the model also delivers dynamic predictions formalizing the qualitative logic of the spiral of silence theory. Equilibrium beliefs about the prevalence of socially appropriate views decrease monotonically in the fraction of individuals who attend to silence and converge to the true belief distribution when attention is universal.

We then test the predictions of this model in a dynamic, unified experimental framework. Given recent discussions about self-censorship on college campuses, we study these questions in the context of UC Berkeley undergraduate students discussing sensitive topics together. Our experiment allows us to test the three central elements of the spiral of silence theory: 1) the prevalence of self-censorship; 2) the role of attention to silence; and 3) the self-reinforcing effects of expression and inference on equilibrium beliefs.

In our experiment, 383 Berkeley undergraduates discuss controversial topics in groups with their peers. The discussion topics include: 1) whether all public schools named after

[3]We use Bayesian updating primarily as a tractable benchmark to generate closed form solutions, in line with similar signaling models in the literature (Braghieri, 2024; Bursztyn, Egorov, and Fiorin, 2020). We relax this assumption in our discussion of the model in Section 2.1 and explore evidence consistent with alternative, salience-based belief formation frameworks in Section 5.3.

controversial historical figures should be renamed, 2) whether race should be explicitly taken into account in college admission processes; 3) whether the death penalty should be abolished; and 4) whether immunizations, such as for Covid and the flu, should be required on campus. Before their own group discussions, students are presented with summaries of previous group discussions, where we randomly vary the salience of silence. Our experimental setup is dynamic, intended to capture how students' perceptions about social norms and expression decisions evolve after observing the stated opinions of others.

We first elicit students' private views and beliefs about the views of other students on these topics in a baseline survey. For example, we ask students whether they personally think all public schools named after controversial historical figures should be renamed and to guess what proportion of other study participants agree with renaming. After the baseline survey, we randomly select a subset of our sample to serve as "First Movers", who participate in their group discussions without any additional information.

We then share a summary of the First Movers' discussions of each topic with the remaining study participants (Second Movers). To causally identify the effect of attention to silence, we randomly vary whether participants receive "Control" or "Treatment" information. The Control information shares the number of First Movers who publicly agreed or disagreed with each topic, while the Treatment information includes the Control information and additionally the number of First Movers who remained silent for each topic. For example, Control participants see a pie chart showing that 7 First Movers publicly agreed with renaming schools and 5 disagreed, while Treatment participants see a pie chart that also includes the 13 First Movers who remained silent on the topic. Importantly, both groups are informed of the total number of First Movers and can thus infer the number of silent participants. The key difference between the Control and Treatment groups is not the amount of information provided but the salience of silence.

After the information treatment, we elicit updated beliefs about the views of other students to measure the immediate effects of increasing the salience of silence. Students then participate in their own discussion groups with either all Control or all Treatment participants to discuss the same topics. In other words, participants who received the same information treatment participate in group discussions together. During these group discussions, we observe public expression, and afterwards, we measure how students update their beliefs about the views of others. This allows us to study dynamics in societies where people pay similar levels of attention to silence.

Results from our experiment first show that the spiral of silence exists in practice. Whether a student privately holds a socially appropriate or inappropriate view significantly predicts whether they self-censor. First Movers who hold socially appropriate views are 34%

more likely to express their opinions publicly compared to those who hold socially inappropriate views. Furthermore, in the status quo, participants appear to pay limited attention to silence. Students in the Control group update their beliefs based on the expressed views and thus overestimate the popularity of socially appropriate views. This, in turn, reinforces self-censorship and exacerbates misperceptions relative to baseline beliefs.

Our results then show that increasing attention to silence can effectively break this spiral of silence. Exogenously drawing attention to silence leads to more accurate beliefs about the views of others and increases the diversity of views expressed publicly. Immediately after receiving the information intervention, Treatment students guess that the socially appropriate view is on average 10% less prevalent compared to Control students. This, in turn, makes Treatment students more willing to express socially inappropriate views publicly. Participants in the Treatment group who privately hold socially inappropriate views are twice as likely to express their opinions relative to their counterparts in the Control group. Corresponding with our model assumptions, there are no significant treatment effects on expression for students who hold socially appropriate views.

After participating in these discussions, the difference in beliefs between Treatment and Control participants further diverges. In the endline survey, Treatment students estimate that socially appropriate views are 15% less prevalent compared to Control students. Over time, Control students increasingly overestimate the popularity of appropriate views while Treatment students' beliefs become more accurate. Taken together, our results show that attention to silence affects both inference and expression, and that these effects amplify and build on each other, resulting in diverging equilibria.

Notably, throughout all stages of the experiment, silence is much more pervasive than lying. Only 2% of participants lie, compared to 60% who stay silent. This empirical pattern is consistent with lying imposing a relatively higher psychological cost than staying silent. When lying is costly in this way, our model shows that silence, rather than lying, becomes a significant driver of self-censorship and misperceptions.

We next provide evidence for social norms and attention to silence as key mechanisms underlying our main experimental results. As a placebo, students also discussed a topic on eliminating daylight savings time which was not socially sensitive. We find that there is no spiral of silence nor significant treatment effects on inference or expression when a topic is socially neutral. In line with *attention* to silence playing a key role, we find that Treatment students have more accurate recall of the number of participants who stayed silent during each Zoom discussion topic relative to Control students. Students are on average also sophisticated in their beliefs about selection bias into silence.

We also extend our main experiment using real data from a public feedback process

conducted by UC Berkeley's Building Name Review Committee. To test whether silence is a larger source of misperceptions compared to other types of information omission, we introduce a third treatment arm ("Omit Agree") that omits the number of public agreeing comments. We find that students in the Omit Agree group make similar inferences and public expression choices as those in the Treatment group, where all expression categories are salient. In contrast, omitting silence increases the perceived prevalence of the socially appropriate view and depresses public expression of the inappropriate view, replicating the results from our main experiment. These results highlight that silence, unlike explicit expressions of agreement or disagreement, is a less salient form of expression. When explicit agreement is excluded, individuals recognize the irregularity and actively pay attention to the omission, whereas the omission of silence is more easily ignored.

As a robustness check, we conduct follow-up rounds of the experiment where we test the sensitivity of our results to specific design choices of the information treatments. In particular, we modify the information given to both Control and Treatment participants so that the instructions clearly state that first movers could agree, disagree, or choose to stay silent. We also remove all data visualizations and instead simply state the summary statistics. These variations make silence more salient to participants in the Control group, generating a weaker information treatment that is potentially closer to realistic settings in which silence is noticeable but not highlighted. Our results remain robust under this revised design: students in the revised Treatment group are significantly more likely to express socially inappropriate views publicly and form more accurate beliefs about the true private opinions of their peers relative to students in the revised Control group. These replication results show that our original experimental results are not driven by the particular way that information is presented to students.

Beyond college campuses, we use nationally representative surveys of public opinion from the American National Election Studies (ANES) from 1974-2020 to test the ecological validity of our results. Under the assumptions of our conceptual framework, the fraction of people who stay silent and the magnitude of misperceptions about the views of others should be positively correlated. Intuitively, overestimation of the popularity of one viewpoint stifles public expression, and silence in turn can lead to less accurate beliefs about others. ANES surveys include questions asking respondents to position themselves, Democratic, and Republican parties on a 7-point scale for various political and socioeconomic issues. We consider non-responses to specific questions to indicate silence and measure misperceptions about the Democratic and Republican parties by comparing the actual and perceived responses of Democratic and Republican respondents respectively. Consistent with the predictions of our model, we find a positive correlation between the share of non-responses and the magnitude

of misperceptions across survey topics. A 1% increase in the share of silence is correlated with a 5% increase in misperceptions about others' views on a topic.

Taken together, our results have important policy implications for how to design and share information from commonly used public feedback processes, such as social media posts, opinion polls, public petitions, or town halls. Such feedback processes typically only report the views that are shared. For example, opinion polls typically share what percentage of respondents "agreed" or "disagreed" with a statement, but rarely report the number of respondents who skipped or refused to answer the question. Our study results illustrate that such information on silence is informative about the true distribution of views and that omitting such details is not only a loss of information but can actually increase bias in beliefs about public opinion.

Related Literature. Our paper speaks directly to the political science literature on the Spiral of Silence, introduced in Noelle-Neumann (1974).[4] Existing studies on the spiral of silence focus on the relationship between perceptions about the opinion climate and expression decisions, using mostly survey evidence (Glynn, Hayes, and Shanahan, 1997; Hayes, Glynn, and Shanahan, 2005; Matthes, 2014; Gonzenbach, 1992; Moreno-Riano, 2002; Yun and Park, 2011). Our paper formalizes these ideas with a model and uses experimental evidence to identify the key role of inattention to silence in perpetuating the dynamic effects between self-censorship and misperceptions. In doing so, we build upon an existing literature that shows evidence separately for specific components of this spiral. For example, Bursztyn, Fujiwara, and Pallais (2017) demonstrate that people report different views in private versus public settings; Braghieri (2024) shows that college students fail to take into account the distortions caused by social image concerns when interpreting public statements, leading to systematic misperceptions; and Bursztyn, González, and Yanagizawa-Drott (2020) establish that correcting these misperceptions influences public behavior.

Our central contribution to this literature is to trace the pathway from self-censorship to misperceptions to behavior change within a single, transparent experimental framework. This direct approach provides a causal and unified test of the mutual reinforcement between self-censorship and misperceptions. In doing so, our study speaks to a growing literature showing that misperceptions about social norms are widespread and that correcting such misperceptions can lead to meaningful changes in behavior (Bursztyn, Egorov, and Fiorin, 2020; Bursztyn, González, and Yanagizawa-Drott, 2020; Bursztyn and Yang, 2021). An open question in this literature is how misperception arise and why they persist. While there are multiple theoretical frameworks to explain the persistence of misperceive social norms, well-

[4]See Glynn, Hayes, and Shanahan (1997); Scheufle and Moy (2000); Matthes, Knoll, and Sikorski (2018) for meta analyses

identified dynamic evidence that isolates specific mechanisms remains limited.[5] Our results show that inattention to silence is an underlying mechanism that can explain the origin and persistence of misperceived social norms. Consequently, our study also causally identifies a method for correcting misperceptions, namely increasing the salience of silence.

Importantly, we focus on the role of silence, extending the existing literature on misperceptions of social norms, where silence has typically not been a studied response option. While theoretically there may be little difference between lying and silence, our results show that in practice silence is both especially prevalent and especially overlooked and is therefore a distinctly important source of misperceptions. These results align with the survey methodology literature, which finds that respondents are generally more likely to opt out of answering than to deliberately provide false information to survey questions that impose high cognitive costs or social desirability bias, making silence a dominant mode of non-expression in practice (Krosnick, 1991; Krosnick and Presser, 2010). As we find in our setting, silence may be more prevalent than lying because it imposes lower psychological costs. For example, Spranca, Minsk, and Baron (1991) documents the existence of omission bias: people judge harmful actions as worse, and often less acceptable, than equally harmful inactions.

The fact that we identify silence as a particularly important source of misperceptions corresponds with existing evidence that inherently less salient cues, such as "no news" or "nothing," are informative but often overlooked, both in the lab (Esponda and Vespa, 2018; Enke, 2020; Jin, Luca, and Martin, 2021) and in finance and marketing (Hirshleifer and Teoh, 2003; Li and Hitt, 2008; Koehler and Mercer, 2009; Giglio and Shue, 2014). Another form of this phenomenon is selection neglect, which has been documented in a range of settings, from the lab (Tversky and Kahneman, 1974; Enke, 2020; Deversi, Ispano, and Schwardmann, 2021; Lifchits et al., 2021; Montero and Sheth, 2021; Lopéz-Pérez, Pintér, and Sánchez-Mangas, 2022; Barron, Huck, and Jehiel, 2024), to finance (Koehler and Mercer, 2009; Jehiel, 2018), to political beliefs (Brundage, Little, and You, 2024). More generally, our results speak to a large behavioral economics literature that conceptualizes attention as limited, thus making attention, and how it is directed and spent, relevant to outcomes beyond just the information available in a given context (Bordalo, Gennaioli, and Shleifer, 2012, 2013, 2022; Chetty, Looney, and Kroft, 2009; Brown, Hossain, and Morgan, 2010; Karlan et al., 2016; Stango and Zinman, 2014; Hanna, Mullainathan, and Schwartzstein, 2014; Taubinsky and Rees-Jones, 2017; Fang et al., 2020; Graeber, 2022; Bordalo et al., 2025).

[5]The theoretical frameworks include signaling motives and reputational concerns from the information supply side (Morris, 2001; Bénabou and Tirole, 2006; Benabou and Tirole, 2011); and motivated reasoning (Bénabou and Tirole, 2016), pluralistic ignorance (Kuran, 1995; Shamir and Shamir, 2000; Bicchieri, 2005; Fernández-Duque, 2022), stereotyping (Bordalo et al., 2016), and confirmation bias (Nickerson, 1998) from the information demand side.

In what follows, we present a conceptual framework in Section 2. In Section 3, we describe the experimental design in more detail. In Section 4, we present the main experimental results on inference, expression, and dynamics. In Section 5, we discuss evidence identifying social norms and attention to silence as key mechanisms driving the experimental results. We discuss the robustness and external validity of our findings in Section 6. Finally, we conclude with Section 7, considering lessons from the study and implications for policy.

2 Motivating Framework

We conceptualize this setting as a two-stage signaling model involving two types of individuals: information senders and receivers. In the first stage, senders decide whether and how to express their opinions publicly, based on their private preferences and their perceptions of receivers' views. Upon observing senders' expression decisions, receivers update their beliefs about the distribution of opinions in the population. Importantly, receivers' belief-updating process is affected not only by the opinions that are expressed but also by the choices of those who remain silent - provided that receivers pay attention to silence. In the second stage, receivers become senders and make their own public expression decisions to a new audience.

Formally, each individual is characterized by a type $\theta_i \in \{A, D\}$ regarding a given policy, where $\theta_i = A$ indicates that individual i agrees with the policy, and $\theta_i = D$ indicates disagreement. We assume that agreement ($\theta_i = A$) is socially appropriate and denote by p the actual fraction of individuals holding this view. For notational simplicity, we distinguish between two roles: senders will be indexed by i and receivers will be indexed by j.

Expression Decisions by Senders:

When considering whether and how to express their opinions publicly, senders' decisions depend critically on their private preferences. Individuals who privately hold the socially appropriate view ($\theta_i = A$) face no risk of social sanction when expressing their opinions.[6] Therefore, truthful expression strictly dominates lying, and expression decisions for this type are governed solely by the intrinsic value of honest expression: $V_i \sim U(\underline{v}, \overline{v})$, where $\underline{v} < 0 < \overline{v}$. These types of senders choose between honest expression and silence, and express their opinions whenever the intrinsic value of expression is positive.

In contrast, individuals who privately hold the socially inappropriate view ($\theta_i = D$) weigh three key factors: (1) the intrinsic value from voicing their opinions honestly, (2)

[6]This assumption is supported empirically by our results and also correspond with Hager et al. (2025), who find that the possibility of punishment is more salient than praise in determining how social norms affect political participation.

disutility from potential social backlash borne by individuals who are inferred to be the socially inappropriate preference type and imposed by those holding the socially appropriate view, and (3) psychological costs from misrepresenting their authentic preferences (lying). Formally, for individuals with $\theta_i = D$, her utility from choosing $e_i \in \{A, 0, D\}$ is:

$$U_i(e_i; \theta_i = D, \lambda_i) = V_i \cdot \mathbf{1}\{e_i = D\} - \chi Pr_i(\theta_j = A) Pr_j(\theta_i = D | e_i, \lambda_j) - C \cdot \mathbf{1}\{e_i = A\}$$

where the expression variable $e_i \in \{A, 0, D\}$ takes the value $e_i = \theta_i$ when the sender expresses their true opinion, and $e_i = 0$ when they choose to remain silent. The parameter $\chi > 0$ captures the *injunctive norm* associated with the topic, or the strength of the underlying normative or moral judgment about a position. $\Pr_i(\theta_j = A)$ is sender i's prior about the fraction of receivers who hold the socially appropriate view, who may sanction perceived dissenters. Meanwhile, receiver j updates his beliefs $\Pr_j(\theta_i = D | e_i, \lambda_j)$ about the sender's type with Bayes rule, conditioning on both the observed expression e_i, and receiver j's attention to silence $\lambda_j \in \{0, 1\}$. Finally, $C > 0$ denotes the psychological costs when misrepresenting one's true preferences.

Intuitively, senders holding the socially inappropriate view trade off honestly speaking up and avoiding potential social backlash. The perceived social sanction costs increase when (a) senders believe that most receivers hold the socially appropriate view (i.e., $\Pr_i(\theta_j = A)$ is high); and (b) receivers are more likely to infer dissent from the sender's expression decisions (i.e., $\Pr_j(\theta_i = D | e_i, \lambda_j)$ is high).

We further make two assumptions related to these two perceived probabilities. First, we assume heterogeneous priors with a common distribution: each sender i' prior π_i about $Pr_i(\theta_j = A)$ is given by $\pi_i \overset{\text{iid}}{\sim} Beta(a, d)$. Second, we assume projection of attention parameters. When forming beliefs about how receiver j will perceive him, sender i projects his own attention parameter onto receiver j, such that $\hat{\lambda_j} = \lambda_i$. That is, $\Pr_j(\theta_i = D | e_i, \lambda_j) = \Pr_j(\theta_i = D | e_i, \lambda_i)$. Intuitively, the sender assumes others behave like themselves regarding attention to silence.

We consider a simplified case in which individuals either fully attend to silence ($\lambda_i = 1$) of completely ignore silence ($\lambda_i = 0$), where the distribution of these two types in the society is given by:

$$\Pr(\lambda_i = 1) = m \quad \text{and} \quad \Pr(\lambda_i = 0) = 1 - m.$$

For senders who are inattentive to silence ($\lambda_i = 0$), their belief that receivers similarly ignore silence implies zero sanction probability when remaining silent: $Pr_j(\theta_i = D | e_i = 0, \lambda_i = 0) = 0$. Conversely, attentive senders ($\lambda_i = 1$) anticipate that receivers infer their types from silence in a Bayesian way. For simplicity, we denote $\rho_i \equiv Pr_j(\theta_i = A | e_i =$

$0, \lambda_i = 1)$ as the probability that receivers attribute agreement to silent senders, with dissent probability $Pr_j(\theta_i = D|e_i = 0, \lambda_i = 1)$ consequently being $1 - \rho_i$.

Senders holding the socially inappropriate view, $\theta_i = D$, may misrepresent their views to avoid social sanction costs. When lying, they incur no social sanction costs but bear an internal cost C. For senders inattentive to silence ($\lambda_i = 0$), silence strictly dominates lying. For attentive senders ($\lambda_i = 1$), lying is suboptimal whenever the internal cost exceeds the expected sanction from silence: $C > \xi$, where $\xi = \chi\mathbb{E}[\pi_i](1 - \rho_i)$. A sufficient condition for silence to strictly dominate lying is $C > \chi$.

When the psychological costs of lying are sufficiently large ($C > \chi$), each sender can either remain silent or truthfully disclose her opinion. Under the maintained assumption $\hat{\lambda}_j = \lambda_i$, the sender's probability of silence is given by

$$\Pr_i\big(e_i = 0 \mid \theta_i = A\big) \;=\; G\big(0\big), \quad \Pr_i\big(e_i = 0 \mid \theta_i = D, \lambda_i = 1\big) \;=\; G\big(\chi\,\rho_i\mathbb{E}[\pi_i]\big),$$

$$\Pr_i\big(e_i = 0 \mid \theta_i = D, \lambda_i = 0\big) \;=\; G\big(\chi\,\mathbb{E}[\pi_i]\big)$$

where $G(\cdot)$ is the cumulative distribution of the intrinsic value of expression V_i, ρ_i denotes sender i's belief (when $\lambda_i = 1$) about the fraction of silent individuals who are type A, and is endogenously determined as the fixed point:

$$\rho_i \;=\; \frac{\mathbb{E}[\pi_i]\,G(0)}{\mathbb{E}[\pi_i]\,G(0) \;+\; (1 - \mathbb{E}[\pi_i])\,G\big(\chi\,\rho_i\mathbb{E}[\pi_i]\big)}$$

Following this expression rule, we denote the signal vector by $\mathcal{S} = (S_A, S_D, S_S)$, where (S_A, S_D, S_S) are the observed numbers of individuals expressing A, expressing D, or remaining silent, respectively.

Inferences by Receivers

We assume that receivers' priors about the fraction of people who hold the socially appropriate view in the society is $\pi_j \overset{\text{iid}}{\sim} Beta(a, d)$. Receivers observe the publicly expressed and unexpressed opinions $\mathcal{S} = (S_A, S_D, S_S)$ and update their beliefs according to these signals. For tractability, we assume Bayesian updating is applied to signals that are attended to.

When $\lambda_j = 0$, receiver j ignores silence and only expressed opinions are used:

$$\pi_{j0}|\mathcal{S} \equiv \pi_j(\lambda_j = 0)|\mathcal{S} \sim Beta(a + S_A,\; d + S_D)$$

When $\lambda_j = 1$, receiver j updates based on both expressed opinions and silence. In this case,

the receiver must simultaneously infer the likelihood that a silent individual is type A:

$$\begin{cases} E[\pi_{j1}|\mathcal{S}] \equiv E[\pi_j(\lambda_j = 1)|\mathcal{S}] = \dfrac{a + S_A + \rho_j S_S}{a + d + S_A + S_D + S_S} \\ \rho_j \;=\; \dfrac{\mathbb{E}[\pi_{j1}|\mathcal{S}]\, G(0)}{\mathbb{E}[\pi_{j1}|\mathcal{S}]\, G(0) \;+\; (1 - \mathbb{E}[\pi_{j1}|\mathcal{S}])\, G\big(\chi\, \rho_j \mathbb{E}[\pi_{j1}|\mathcal{S}]\big)} \end{cases}$$

Proposition 1. *Assuming uninformative priors, and when the cost of lying is sufficiently large ($C > \chi$), the posterior mean belief about the prevalence of the socially appropriate view for a receiver who attends to silence is strictly lower than that of a receiver who ignores silence:* $\mathbb{E}[\pi_{j1}|\mathcal{S}] < \mathbb{E}[\pi_{j0}|\mathcal{S}]$.

Intuitively, receivers who attend to silence infer that some individuals who remain silent are likely to disagree with the socially appropriate view. Therefore, they form a lower posterior belief about the prevalence of the socially appropriate view compared to receivers who ignore silence. Proof of Proposition1 is detailed in Appendix A.1.

In the second stage, receivers become senders and make their own public expression decisions:

Proposition 2. *Given expressed signals $\mathcal{S}$, among individuals with $\theta_j = D$, those who attend to silence are more likely to publicly express D:* $Pr_j(e_j = D|\theta_j = D, \lambda_j = 1) > Pr_j(e_j = D|\theta_j = D, \lambda_j = 0)$.

Proposition 2 follows directly from the inequality $\mathbb{E}[\pi_{j1}|\mathcal{S}] < \mathbb{E}[\pi_{j0}|\mathcal{S}]$, which implies that $G(\chi\rho_i\mathbb{E}[\pi_{j1}|\mathcal{S}]) < G(\chi\mathbb{E}[\pi_{j0}|\mathcal{S}])$. Intuitively, among receivers who hold the socially inappropriate view ($\theta_j = D$), those who pay attention to silence infer that the socially appropriate view A is less prevalent and anticipate weaker social sanctions for expressing D. Additionally, receivers who pay attention to silence believe others also attend to silence and anticipant greater social sanction costs from staying silent. These two effects collectively make them more likely to publicly express their views in the second stage.

In the experiment, we test Propositions 1 and 2 by exogenously varying whether participants' attention is drawn to those who stayed silent in a public discussion, hence experimentally increasing λ.

Dynamics and Equilibrium

We further extend this simple two-stage model to capture the dynamics in this expression and belief-updating process. In each period t, the observed signal vector $\mathcal{S}_t$ is generated according to the underlying distribution of opinions in the population and the expression rules specified above. Receivers use these signals to update their beliefs.

In equilibrium, with the distribution of the attention parameter characterized by m, we denote the equilibrium beliefs as π_1^*, π_0^* and equilibrium expressions as S_A^*, S_D^*, S_S^*.

Proposition 3. *Define the average equilibrium belief in the population as*

$$\mathbb{E}[\pi^*] = (1-m)\mathbb{E}[\pi_0^*] + m\mathbb{E}[\pi_1^*].$$

where $\mathbb{E}[\pi_0^*]$ *and* $\mathbb{E}[\pi_1^*]$ *denote the average equilibrium beliefs of individuals who ignore and attend to silence respectively. Under mild regularity conditions ensuring sufficient heterogeneity in intrinsic expression values, the following equilibrium results hold:*

1. $\mathbb{E}[\pi_0^*] > \mathbb{E}[\pi_1^*]$ *for all* $m \in [0,1]$.
2. $\mathbb{E}[\pi^*] = \mathbb{E}[\pi_1^*] = p$ *when* $m = 1$.
3. $\mathbb{E}[\pi_0^*]$, $\mathbb{E}[\pi_1^*]$, *and* $\mathbb{E}[\pi^*]$ *decrease monotonically in* m.

Intuitively, the average equilibrium belief $\mathbb{E}[\pi]^*$ depends on the fraction m of individuals who pay attention to silence. The existence of individuals who do not pay attention to silence leads the population to overestimate the prevalence of the socially appropriate view. As m increases, this overestimation weakens, and the perceived distribution of beliefs moves closer to the true distribution.

Distinguishing the individual-level attention parameter λ_i from the population share m allows us to clarify the mechanism in two ways. First, holding m fixed, we can characterize and compare the equilibrium beliefs of individuals who attend to silence and those who do not, rather than treating all agents symmetrically. Second, we can examine how equilibrium beliefs vary with m, highlighting how the prevalence of attention to silence shapes aggregate belief dynamics in the population.

Specifically, part (1) if Proposition 3 states that, for any given m, individuals who ignore silence systematically hold higher equilibrium beliefs than those who pay attention to silence. Part (3) implies that increasing the fraction m of silence-attentive individuals lowers equilibrium beliefs for both types of receivers. Finally, part (2) establishes that when everyone pays full attention to silence ($m = 1$), equilibrium beliefs are correct on average: the posterior mean belief equals the true prevalence p. A detailed proof of Proposition 3 can be found in Appendix A.2.

2.1 Discussion of framework

In our model, we distinguish conceptually between two distinct mechanisms through which social norms can affect expression decisions. The parameter $\chi > 0$ captures the *injunctive norm* associated with the topic - that is, the strength of the underlying normative or moral judgment about a position. On the other hand, $\Pr_i(\theta_j = A)$ captures how often an individual

expects to encounter sanctioning when expressing socially inappropriate views, representing the *descriptive norm*. This second term captures the component of social sanctions which depends explicitly on the believed popularity of views. While in some real-world contexts these two forces may be correlated (e.g. stronger norms can coincide with wider social consensus), in our model we treat them as separable parameters, which can help to disentangle these two conceptually distinct mechanisms of social sanctions. For example, some social norms are believed to be determined, at least in part, by changes in policy or legislation.[7] In our model, the effects of a new legislation on social norms would be captured by χ.[8]

We also assume that if individuals pay attention to silence, they update from it as a Bayesian. In reality, however, individuals may not update in a Bayesian way, especially if attention and beliefs are heterogeneous or multi-modal across individuals. We use Bayesian updating primarily as a tractable benchmark to close the model and to derive clear comparative statics, following the standard approach in closely related work such as Braghieri (2024), Bursztyn, Egorov, and Fiorin (2020) and other signaling models. However, to test the sensitivity of our results to this modeling assumption, we conduct a series of simulation exercises that compare the dynamics of beliefs under alternative updating rules. Figure A.1 reports the simulated belief dynamics using different updating rules: Bayesian updating, a mechanical 50/50 updating rule that treats silence as "50% Agree, 50% Disagree," projection of one's prior beliefs onto those who stay silent, and projection of one's own private type onto silence. Across all updating rules we consider, the perceived popularity of the socially appropriate view exhibits a monotonic comparative static in the share of individuals who pay attention to silence. When a larger fraction of individuals pay attention to silence (higher m), the society as a whole infers a lower share of agreement $E[\pi]$. This monotonic relationship, which is the key prediction that we test experimentally, does not depend on a specific belief-updating rule.

[7]For example, the U.S. Supreme Court ruling in Obergefell v. Hodges that legalized same-sex marriage coincided with a significant increase in public support for same-sex marriage. For a more detailed discussion of this example, see `https://duncanwebb.substack.com/p/how-the-world-became-more-accepting?r=bc4om&utm_campaign=post&utm_medium=web&triedRedirect=true`

[8]The model could also be extended to allow for persuasion to be another motivation for public expression. For example, for someone who holds the socially inappropriate view, their utility of expression would be represented by $V_i - (\chi - \alpha) \cdot E[\pi]$, where α denotes the persuasion motive, or the likelihood that expressing a view publicly changes the views of others. Empirically, we find that $\chi > \alpha$ because the willingness of expression decreases in the perceived popularity of the socially appropriate view $E[\pi]$. Therefore, in our context, the social sanction costs outweigh the persuasion motive on average.

3 Experimental Design

To test the spiral of silence hypothesis, we conduct an experiment with UC Berkeley undergraduate students. We first solicit participants' private views on a set of socially sensitive topics and then provide all participants with an opportunity to express their views publicly. Capturing both private and public opinion allows us to determine the prevalence of self-censorship and whether socially appropriate views are expressed more often in public. To test whether attention to silence affects beliefs about others and public expression over time, we experimentally vary the salience of silence when sharing information about discussions by previous groups. We then assign participants to public discussion groups with other participants who received the same information treatment to test the dynamic, self-reinforcing effects predicted by our conceptual framework.

3.1 Study Procedures

3.1.1 Baseline Survey

All participants first complete a baseline survey about their private views on 10 potentially sensitive socioeconomic and political topics. Participants are then asked to guess the views of other students on the same set of topics. Specifically, for private views, we ask participants to choose either "agree" or "disagree" for each statement.[9] For guesses about the views of other students, we ask *"Among all other Berkeley students who participate in this study, what percentage do you think will privately answer 'agree' and 'disagree' [with this statement]?"*. Survey responses about the beliefs of other students' views are incentivized for accuracy using a binarized scoring rule.[10] We also collect demographic information, including gender, ethnicity, school year, major, and self-identified political affiliation.

3.1.2 Group Discussion Sessions

Discussion Procedure: All respondents then participate in Zoom sessions of approximately 8-13 participants, where they have the opportunity to express their views publicly to other students on a subset of topics. To make the Zoom session feel more like public discourse with peers, participants are asked to keep their camera on throughout the session and to display

[9]Participants can also choose to skip any question if they do not want to provide their private views. In practice, only 1% of participants skipped questions about their private beliefs in the baseline survey.

[10]Danz, Vesterlund, and Wilson (2022) show that simplifying the instructions of BSR increases the accuracy of belief elicitation. In light of this, we include a non-quantitative description about incentives in the survey instructions and inform participants that "you will maximize your chance of earning the bonus for each statement if you report your beliefs as accurately as possible". We include a link to the quantitative details that participants can access if interested. The scripts are similar to Burdea and Woon (2022) and can be found in Appendix F.

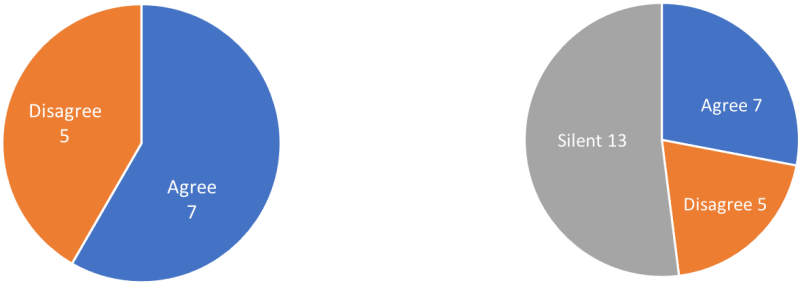
their first name and the first initial of their last name. To receive study compensation, students just need to be present during the Zoom discussion, but they are not paid based on whether or how often they speak. Each Zoom session is facilitated by a trained moderator using a standardized script.

Within each Zoom session, the same four potentially socially sensitive topics are discussed: Renaming Schools, Affirmative Action, Death Penalty and Immunizations. Each topic is first briefly introduced by the moderator, who then reads a statement on the topic. For example, for the renaming schools topic, the moderator reads the statement *"All public schools named after controversial historical figures, including former Presidents George Washington, Thomas Jefferson, and Abraham Lincoln, should be renamed."* All participants then have 90 seconds to decide whether they would like to share their views on the topic with other students in the Zoom room. Participants first message the moderator privately indicating whether they "agree" or "disagree" with the statement if they want to share their views publicly. Students who do not wish to share their views simply do not message the moderator. The moderator then calls upon all of the participants who volunteered in a random order to share their views. All participants who indicate they want to share their views are called upon to do so. To shut down potential dynamic effects within each session, the order in which participants are called upon is random, and participants speak only when called. A detailed script for the Zoom sessions can be found in Appendix G.

First Movers: A small subset of participants are randomly assigned to be First Movers, who participate in Zoom sessions with other First Movers without any additional information. In total we hosted 4 Zoom Sessions with First Movers, each with 12-13 participants.

Second Movers: We then share a summary (anonymized) of the views expressed publicly by the First Movers with the remaining participants (Second Movers). Importantly, we randomly vary whether Second Movers receive "Control" or "Treatment" information about the public views of the First Movers. Specifically, Control participants receive information about the share of First Movers who publicly agreed or disagreed with each topic. For example, we inform Control participants that *"25 Berkeley students like you discussed over Zoom if they agreed with the following statement ..... Here is a summary of their Zoom discussion"* along with a pie chart showing 7 "Agree" and 5 "Disagree" (see the left panel of Figure 1). Treatment participants receive the same instructions and information as Control participants plus additional information on the number of First Movers who remained silent on each topic. Using the same example as above, Treatment students see a pie chart showing 7 "Agree", 5 "Disagree" and 13 "Silent" (see the right panel of Figure 1). Note that both groups are provided with sufficient information (total number of students in the previous

sessions) to determine the number of participants who stayed silent. The key difference between the Control and Treatment information is not the amount of information provided but the *salience* of silence.



Control group Treatment group

Figure 1: Information Treatments

Note: This figure shows an example of the information interventions shared with Control and Treatment group participants respectively, summarizing the discussions of the First Movers of the Renaming Schools topic. The information intervention was shared during the midline survey, immediately before participants participated in Zoom discussions together.

The information treatment is embedded in a midline survey that is sent to all Second Movers approximately 12 hours before their respective Zoom sessions. To measure the immediate effects of the information treatment on beliefs, we also re-elicit participants' beliefs about the views of other students on this set of socially sensitive topics.

For the Second Movers' Zoom sessions, Control participants are only eligible to sign up for Control Zoom sessions and Treatment participants are only eligible to sign up for Treatment Zoom sessions. In other words, students participate in Zoom sessions with only other students who received the same information. This allows us to study dynamics in societies where people pay similar levels of attention to silence. The protocol for the Zoom sessions for all Second Movers, including Control and Treatment participants, is otherwise identical to that of the First Movers.

In total, we hosted 36 Zoom sessions with Second Movers, 18 Control and 18 Treatment sessions. Each Zoom session was attended by on average 9 participants, with the smallest session containing 8 attendees and the largest session containing 12 attendees. All participants only participated in one Zoom session. The Zoom sessions each lasted 45 minutes and took place within one to two weeks of the participant completing the baseline survey. The day of week and time of day of each Zoom session was balanced across Control and Treatment sessions so that students' scheduling availability would not be a potential confound.

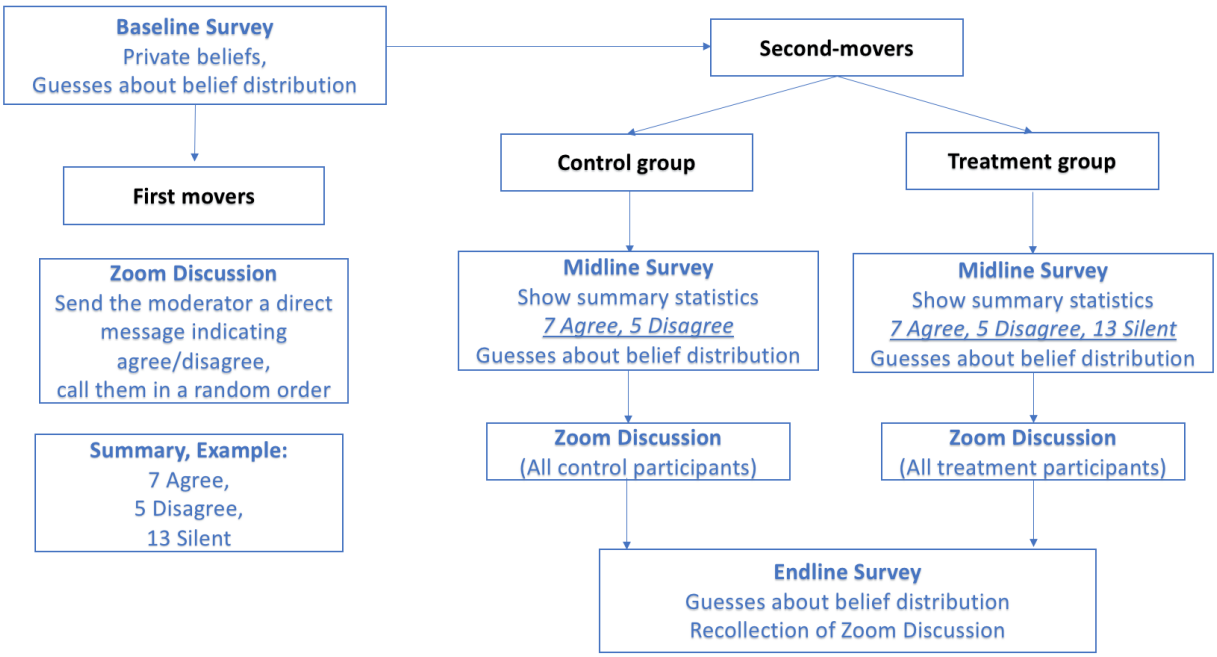


Figure 2: Experimental Design

3.1.3 Endline Survey

At the end of each Zoom session, participants complete an endline survey where they again provide their guesses about the distribution of other students' views. This endline survey allows us to measure how participants update their beliefs after participating in public discourse with other Control/Treatment peers respectively. To assess whether attention to silence is a mechanism, we also elicit participants' recollections of their Zoom discussions, including how many students expressed agreement, disagreement, or stayed silent on each topic. We also ask participants to guess how students who stayed silent would have privately agreed or disagreed with each statement to measure beliefs about selection into silence. Figure 2 illustrates our experimental design in full and detailed survey instruments can be found in Appendix F.

3.2 Topics

We selected the set of socially sensitive topics for our experiment based on current events and opinion articles in popular news outlets and on social media as well as from focus groups with separate samples of college students from Prolific.[11] For our main experiment, we

[11]For example, the topic on Renaming Schools was based on a decision by the San Francisco School Board in Jan 2021 to rename 44 school sites that honored controversial historical figures. The decision was later unanimously reversed in Apr 2021 after public outcry. The topic on Affirmative Action was based on the Students for Fair Admissions vs. Harvard and UNC lawsuit. These issues were covered extensively in the media, such as in the Atlantic, the Guardian, the New York Times, and the Wall Street Journal.

covered the following five topics in Zoom sessions, which included one placebo topic that is not socially sensitive:

- Renaming Schools: All public schools named after controversial historical figures, including former Presidents George Washington, Thomas Jefferson and Abraham Lincoln, should be renamed.
- Affirmative Action: If Proposition 209 was repealed, universities in the UC system should adopt extensive affirmative action policies that explicitly take into account race in the admission process.
- Death Penalty: The U.S. should abolish the death penalty.
- Immunizations: Immunizations, such as for Covid and flu, should be required on Berkeley's campus.
- Daylight Savings Time "DST" (Placebo): Daylight savings time should be eliminated.

Recall that in the baseline survey we elicited both the private beliefs and the beliefs of other students' views on each of these topics. Figure 3 shows the actual percentage of students who privately choose "agree" for each statement, along with students' guesses about the percentage of students who choose "agree." These results show that misperceptions are widespread at baseline: there exist statistically significant differences between actual and perceived belief distributions for all five topics. For example, on average participants estimated that 56% of their fellow Berkeley student respondents would privately agree with Renaming Schools, when in reality only 39% of respondents privately agreed with this statement. In a similar vein, students estimate that 53% of other students would agree with Affirmative Action, while the actual share of agreement is significantly lower at 46%.

To measure the social appropriateness of each of these statements, we follow Krupka and Weber (2013) and invite Berkeley students to rate the social appropriateness of these statements using simple coordination games. Students are asked to rate whether each statement is "very inappropriate," "somewhat inappropriate," "neither appropriate nor inappropriate," "somewhat appropriate," or "very appropriate" and are incentivized to match their response with the modal response.[12] We then use these survey responses to determine which direction of each statement is socially appropriate, pooling together "somewhat appropriate" and "very appropriate" and likewise "somewhat inappropriate" and "very inappropriate." Empirically, the Krupka-Weber measure provides a natural proxy for the parameter χ from our conceptual framework.

[12]That is, participants were told that they would earn higher payment if their ratings are the same as the most common answer from other survey respondents.

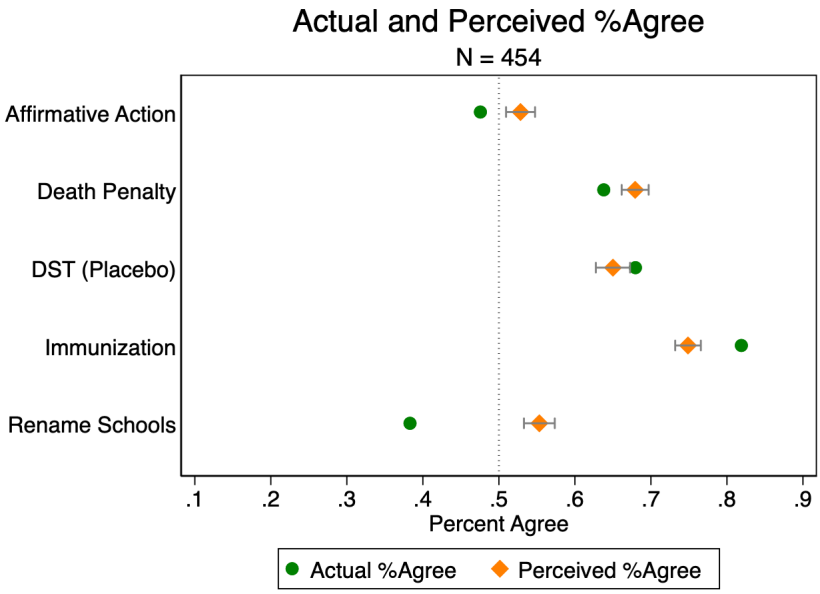


Figure 3: Baseline Actual and Perceived Fraction of Student Agreement

Notes: In this figure, the green dot shows the actual share of agreement among all private views expressed by participants in the baseline survey. The orange diamond represents participants' guesses, incentivized for accuracy, of the share of agreement among all other study participants for each topic in the baseline survey, with 95% confidence intervals displayed.

Elicitation of social appropriateness was done in a separate survey with First Movers after they completed all other parts of the baseline survey, to avoid biasing their other responses. Table B.1 presents the distribution of student responses, with bold numbers representing the modal response for each topic. Among the four sensitive topics discussed in the Zoom session, "somewhat appropriate" is the modal response in the Krupka-Weber measure, suggesting that these issues are viewed as moderately, but not strongly, normatively charged. Nearly half of the participants consider the statement about eliminating daylight savings time to be neutral.

Discussion on Topic Selection for Zoom Sessions:

Apart from the five topics that are discussed in Zoom sessions, we also include some other topics in the baseline survey to reduce experimenter demand effects. A detailed description of these topics can be found in Appendix B.1.

Due to time constraints and feasibility, we could only include a subset of the topics in the baseline survey in the Zoom discussions. As our study is focused on testing the spiral of silence, we select topics where participants overestimate the popularity of the socially appropriate view. Thus, our study results should not be seen as representative of the universe of potentially socially sensitive topics, but rather is specific to topics where people misperceive the popularity of the socially appropriate view. Based on these criteria, we chose the topics on Affirmative Action, Death Penalty, and Renaming Schools to be discussed in the Zoom

sessions. We also included Immunizations as a case where the socially appropriate view is underestimated, and Daylight Saving Time as a placebo. In Appendix B.3, we discuss in more detail how we selected which topics from the baseline survey to include in the Zoom discussion sessions.

3.3 Sampling and Attrition

We recruited our sample through UC Berkeley's Experimental Social Science Laboratory (Xlab), whose participant pool includes current UC Berkeley students and recent graduates who voluntarily participate in research studies. In total, we recruited 454 Berkeley students over five rounds from late February to mid April in 2023 through Xlab's online recruitment system SONA. In terms of demographics, 70% of our participants are female, 54% are Asian and 21% are White. The majority of participants are in their junior or senior year and identify themselves as "Liberal" or "Slightly liberal" (see Panel A in Table C.2).

We recruited in rounds to maintain consistency in the time lag between the baseline survey and subsequent Zoom sessions across all participants and to offer participants more scheduling choices when registering for Zoom sessions. In each round, we recruited approximately 100 participants.[13] Participants first completed a baseline survey and then signed up for a Zoom discussion session scheduled for the following week. Participants could select from four designated Zoom sessions based on their treatment assignment. All sessions were scheduled between 3-5pm, Mondays through Thursdays, to eliminate potential confounds that might arise due to different availability on various weekdays or times. Once a student participated in one of the study rounds they were ineligible to participate in subsequent rounds.

Among participants recruited in the first round, a subset of students were randomly assigned to the "First Movers" group and the remaining students to either the "Control" or "Treatment" group. Participants recruited in the subsequent rounds were randomly assigned to either the Control or Treatment groups. In each round, an equal number of students were assigned to the Control and Treatment groups, so recruitment round is balanced across groups. In total, 56 participants were randomly assigned to the First Movers group, 200 to the Control Group, and 198 to the Treatment group respectively. Assignment of participants to First Movers, Control, and Treatment groups was balanced across observable characteristics and baseline views and beliefs (see Panel B in Table C.2).

In terms of attrition, 454 participants completed the baseline survey, including 56 First Movers and 398 Second Movers. Among the 398 Second Movers who completed the baseline survey, 353 (90%) signed up for a Zoom session and 333 (84%) actually attended their

[13]Specifically, we recruited 95, 101, 63, 88 and 107 participants during the weeks of February 14, February 28, March 7, March 14 and March 28 respectively.

scheduled session. All participants who attended Zoom sessions completed the endline survey. In total, our final sample consists of 383 students who completed all components of the study, including 50 first movers and 333 second movers, of which 166 were assigned to the Control group and 167 were assigned to the Treatment group. Attrition was balanced across treatment assignment, observable characteristics, and baseline views and beliefs (Table C.3). Furthermore, students who dropped out in the Treatment versus Control groups do not differ systematically in observed characteristics including demographics, private beliefs, and baseline guesses about the belief distribution (Table C.4).

3.4 Discussion: From Model to Experiment

Our experimental design allows us to answer three main research questions. In our conceptual framework, we assumed that people with socially inappropriate views face social sanction costs to publicly dissenting and are thus more likely to self-censor. This is the starting assumption of the spiral of silence theory, and we test it by comparing the views expressed privately in the baseline survey versus publicly in Zoom sessions among First Movers.

Second, by experimentally varying the salience of silence and comparing the perceived belief distributions between Treatment and Control groups in the midline survey, we identify the causal effect of increasing attention to silence on perceived social norms. The key parameter in our conceptual framework is λ, the level of attention to silence. In our experiment, we exogenously increase λ to measure its impact on beliefs about others and public expression decisions. This is a direct test of Propositions 1 and 2.

Finally, by comparing the public discourse that occurs in Control and Treatment Zoom sessions, and also the perceived belief distributions in the endline survey, we test the dynamic effects predicted by the spiral of silence theory. Proposition 3 predicts that different levels of attention to silence produce divergent perceived social norms in equilibrium through the self-reinforcing effects of self-censorship and inattention to silence. By grouping Control students and Treatment students together, we exogenously vary the distribution of attentive types, m across different "societies", testing the predictions on dynamics generated by our model.[14]

4 Experimental Results

As described in Section 3.2, we covered four socially sensitive topics in the Zoom discussions: Renaming Schools, Affirmative Action, Death Penalty, and Immunizations. Through-

[14]Note that we pre-registered the experimental design, treatment assignments, and target sample size, but not a comprehensive analysis plan. The heterogeneous treatment effects analyzed in the paper are motivated by the theoretical framework rather than introduced ad hoc.

out our analysis, we standardize responses so that the "Agree" stance represents the socially appropriate view. In the following sections, we combine these four topics when presenting the main results. Results by topic can be found in Appendix C.

4.1 First movers

The starting assumption in the spiral of silence theory is that individuals will be more inclined to stay silent if they believe that their opinion is socially inappropriate. We first provide some suggestive evidence that this is indeed the case by comparing the public expressions and private beliefs of First Movers.

For First Mover i on topic j, we estimate the following regression:

$$Express_{i,j} = \beta_1 Agree_{i,j} + X_i + \alpha_j + \varepsilon_{i,j} \tag{1}$$

where $Express_{i,j}$ is a dummy variable that is equal to one if participant i publicly expresses their opinion on topic j in their Zoom session; $Agree_{i,j}$ indicates whether participant i privately agrees with the statement on topic j; α_j are topic fixed effects, and X_i is a vector of individual control variables including gender, race and ethnicity, year in school, major, political affiliation, and baseline guesses about the prevalence of the socially appropriate view ($\%Agree$). To account for potential correlations in the expression residual within subjects, we cluster standard errors at the individual level. The spiral of silence model assumes that β_1 will be positive.

Our results from the First Movers demonstrate that participants who hold socially appropriate views are approximately 34% (16 percentage points) more likely to express their opinions publicly, suggesting that silence is indeed selected by private beliefs interacted with perceived social norms (see Table 1). Our results remain similar when using either an OLS or Logit model to estimate Equation 1. We caution against interpreting the positive coefficients as causal evidence because there could exist omitted variables that are correlated with both private beliefs and public expression decisions. Nevertheless, the expression decisions by First Movers suggest that silence is selected based on private beliefs.

Recall that the primary objective of having First Movers in our experiment is to generate public expression signals to share with receivers, or Second Movers, who form the main sample of interest for our experiment. For our information treatments, we used data from two randomly selected First Mover Zoom sessions, which included 25 First Movers combined. We did not use data from all First Mover Zoom sessions to avoid overly-shifting Second Mover beliefs in the midline survey. Table C.5 displays the summary statistics from the selected First

Table 1: Public Expression Decisions by First Movers

	(1) Express = 1	(2) Express = 1	(3) Express = 1	(4) Express = 1
Panel A: OLS				
Private Agree	0.153*	0.152*	0.159*	0.161*
	(0.0796)	(0.0798)	(0.0796)	(0.0833)
Panel B: Logit				
Private Agree	0.151**	0.150**	0.157**	0.162**
	(0.0761)	(0.0761)	(0.0753)	(0.0754)
Topic FE	✓	✓	✓	✓
Baseline guesses		✓	✓	✓
Session FE			✓	✓
Ind Controls				✓
Mean	0.440	0.440	0.440	0.440
SD	0.498	0.498	0.498	0.498
IDs	50	50	50	50
Obs	200	200	200	200

Notes: This table reports the public expression decisions by First Movers, pooling together the four socially sensitive topics discussed in Zoom sessions (Affirmative Action, Death Penalty, Immunizations, and Renaming Schools). $Express = 1$ if individual i truthfully express their opinions on topic j. Individual controls include gender, race and ethnicity, year in school, major, and self-reported political ideology. Standard errors are clustered at the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Mover Zoom sessions that were presented to the Control and Treatment groups.[15] Similar to Figure 1, we showed different pie charts excluding or including the number of First Movers who remained silent for the Control and Treatment groups respectively.

4.2 The Effects of Attention to Silence on Inference

Our experiment allows us to test the causal effects of increasing attention to silence on participants' inference about true public opinion. Our model predicts that increasing attention to silence will lead to lower estimates about the popularity of socially appropriate views. Specifically, we estimate:

$$\%\tilde{Agree}^{M}_{i,j} = \beta_1 Treat_i + \beta_2 Agree_{i,j} + \beta_3 \%\tilde{Agree}^{B}_{i,j} + \alpha_j + X_i + \varepsilon_{i,j} \quad (2)$$

where $\%\tilde{Agree}^{M}_{i,j}$ is respondent i's beliefs at midline about the fraction of participants that privately agree with statement j, after seeing the information treatment; $Treat_i$ is a dummy variable indicating whether i is assigned to the Treatment or the Control group; $Agree_{i,j}$ is i's private beliefs on topic j; $\%\tilde{Agree}^{B}_{i,j}$ is i's baseline guesses about the share of agreement;

[15]Table C.6 report the summary statistics if we used data from all First Movers Zoom sessions. "%Agree", "%Disagree" and "%Silent" are similar as in Table C.5.

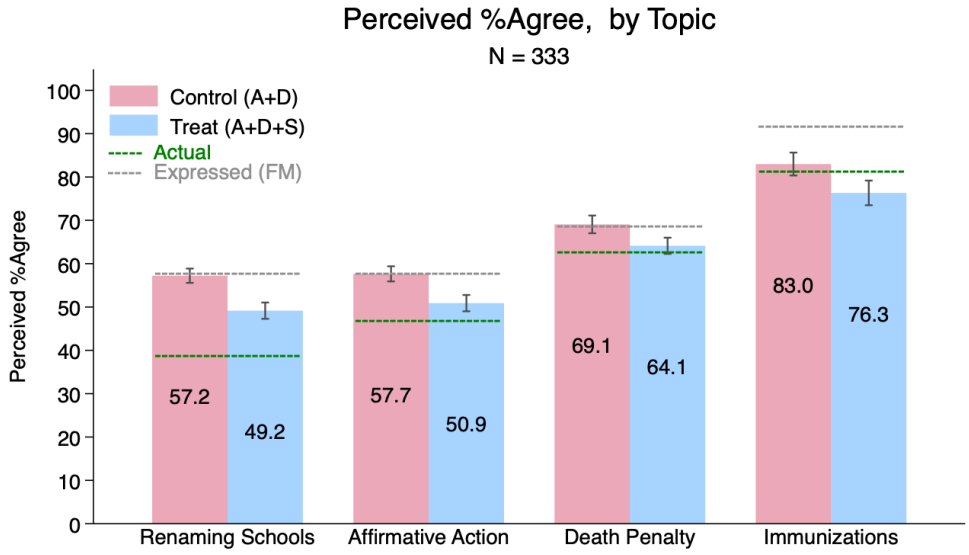


Figure 4: Treatment Effects on Inference at Midline

Note: This graph depicts the perceived share of agreement for each topic elicited in the midline survey. The green dashed line shows the actual share of agreement reported in the private baseline survey for all participants (first and second movers combined). The gray dashed line displays the expressed share of agreement from First Movers that participants are shown as part of both the Control and Treatment interventions. 95% confidence intervals displayed.

α_j are topic fixed effects; and X_i is a vector of individual control variables including gender, race and ethnicity, year in school, major, and political affiliation. Our model predicts that drawing attention to silence will decrease the perceived popularity of socially appropriate views, or that β_1 will be negative.

Results from our midline survey indicate that participants in the Treatment group on average perceive the popularity of the socially appropriate view to be lower compared to the average beliefs of the Control group (see Figure 4). In contrast, the Control group's beliefs move closer to the views publicly expressed by the First Movers.

The combined results from all four topics discussed in the Zoom sessions are presented in Columns 1-3 in Table 2, with Column 3 being our preferred specification corresponding to Equation 2. On average, participants in the Treatment group guess that the socially appropriate view is 10% (6.7 percentage points) less prevalent than the Control group. Additionally, we observe that participants who privately agree with a statement tend to perceive their own views as more popular than those who privately disagree with the statement. The positive coefficients of $\%\tilde{Agree}^B$ suggest positive correlations between the perceived belief distributions elicited across different survey rounds.

Overall, these results show that increasing attention to silence leads to lower estimates of the prevalence of socially appropriate views, as predicted by Proposition 1 in our conceptual framework.

Table 2: Treatment Effects on Inference at Midline and Endline

	Midline Guesses: $\%Agree$			Endline Guesses: $\%Agree$		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	-6.603***	-6.516***	-6.703***	-10.63***	-10.47***	-10.40***
	(0.970)	(0.930)	(0.958)	(1.490)	(1.481)	(1.162)
Private Agree		3.860***	3.198***		3.664***	3.251***
		(0.858)	(0.850)		(0.926)	(0.963)
$\%\tilde{Agree}^B$		0.115***	0.116***		0.110***	0.105***
		(0.0241)	(0.0250)		(0.0270)	(0.0246)
SUEST $\hat{\beta}_1^M = \hat{\beta}_1^E$				10.54	10.55	8.26
				(0.0012)	(0.0012)	(0.0040)
Mean	66.73	66.73	66.73	70.45	70.45	70.45
SD	17.11	17.11	17.11	18.00	18.00	18.00
Topic FE	✓	✓	✓	✓	✓	✓
Baseline guesses		✓	✓		✓	✓
Session Controls						✓
Ind Controls			✓			✓
IDs	333	333	333	333	333	333
Obs	1332	1332	1332	1332	1332	1332

Notes: The outcome variables are the perceived share of agreement among all other study participants ($\%Agree$) elicited in the midline survey and the endline survey for all second movers. Combined results from the four socially sensitive topics (Affirmative Action, Death Penalty, Immunizations, and Renaming Schools) are reported. Standard errors are clustered at the individual level in columns 1-3, and clustered at the session level in columns 4-6. Individual controls include gender, race and ethnicity, year in school, major, and self-reported political affiliation. SUEST reports results from seemingly unrelated regressions testing whether the treatment coefficient is the same for midline and endline guesses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.3 Self-Reinforcing Effects

A key prediction of the *spiral* of silence theory is that the combination of the patterns established above – namely that 1) socially inappropriate views are less likely to be expressed publicly and 2) inattentive individuals overly update their beliefs after observing public discourse – leads to dynamic, self-reinforcing effects. In our experiment, we test for these dynamic patterns by separating Control and Treatment participants into different public forums and then comparing their public expression and subsequent inference. Specifically, our results from Section 4.2 on inference show that exogenously varying the level of attention paid to silence leads to different beliefs about public opinion. Our model then predicts that different beliefs about public opinion in turn lead to different expression decisions. Over time, public forums formed with more individuals who pay attention to silence (Treatment) will diverge from forums where fewer individuals pay attention to silence (Control), resulting in different perceived social norms and public expressions.

We test for the existence of these dynamic patterns using expression decisions from the Zoom sessions with Second Movers as well as our endline survey. First, we test whether individuals who privately hold socially inappropriate views are more likely to express those views publicly in the Treatment group relative to the Control group. To do so, we estimate

the following regression for individual i who privately disagrees with statement j:

$$PublicDisagree_{i,j} = \beta_1 Treat_i + X_i + \theta_k + \varepsilon_{i,j} \tag{3}$$

Second, we then estimate the following regression to measure the causal effects of increasing attention to silence on endline guesses about the belief distribution:

$$\%\tilde{Agree}^{E}_{i,j} = \beta_1 Treat_i + \beta_2 Agree_{i,j} + \beta_3 \%\tilde{Agree}^{B}_{i,j} + X_i + \theta_k + \varepsilon_{i,j} \tag{4}$$

where the outcome variables are $PublicDisagree_{i,j}$, a dummy variable that is equal to one if participant i truthfully expresses their disagreement with topic j; and $\%\tilde{Agree}^{E}_{i,j}$, participant i's endline guesses about the fraction of students who privately agree with statement j. $Treat_i$, $Agree_{i,j}$, $\%\tilde{Agree}^{B}_{i,j}$ and X_i are defined as above; and θ_k are session-level control variables, including time and week of the Zoom session, group size, and moderator fixed effects. Standard errors are clustered at the Zoom session level in both Equations 3 and 4.

Our results indicate that increasing attention to silence significantly increases the likelihood that an individual publicly expresses a socially inappropriate view. For our preferred specification (Column 5 in Table 3), which corresponds to Equation 3, participants in the Treatment group who privately disagree with the socially appropriate view are more than twice as likely to publicly express their opinions, a 18.7 percentage point increase, holding other factors constant, relative to the Control group. Note that there are no significant treatment effects on expression for those who privately agree with the socially appropriate view, corresponding with the assumptions made in our conceptual framework. These results are robust across various specifications, including using a OLS or Logit model and using different sets of control variables. Table 3 pools results across all four Zoom topics to maximize power. Results by topic can be found in Table C.7.

As the degree of belief updating at midline is endogenous, we also instrument for midline beliefs using treatment assignment to isolate the effects of increasing attention to silence on public expression. We use 2SLS to estimate:

$$PublicDisagree_{i,j} = \beta_1 \%\tilde{Agree}^{M}_{i,j} + X_i + \theta_k + \varepsilon_{i,j} \tag{5}$$

where we use $Treat_i$ as an instrumental variable for $\%\tilde{Agree}^{M}_{i,j}$ and the first-stage is estimated by $\%\tilde{Agree}^{M}_{i,j} = \delta_1 Treat_i + X_i + \theta_k + \varepsilon_{i,j}$. Our first stage and reduced-form results are reported in Table C.8. The 2SLS results are reported in Table C.9 and indicate that on average a 1 percentage point decrease in the perceived share of agreement in the midline survey corresponds to respondents being 15% (2.5 percentage points) more likely to express

Table 3: Treatment Effects on Public Expression

	OLS (Express = 1)				Logit (Express = 1)
	(1)	(2)	(3)	(4)	(5)
Panel A: Privately Disagree					
Treat	0.160***	0.161***	0.168***	0.181***	0.187***
	(0.0320)	(0.0320)	(0.0266)	(0.0309)	(0.0342)
Mean	0.164	0.164	0.164	0.164	0.164
SD	0.371	0.371	0.371	0.371	0.371
IDs	278	278	278	278	278
Obs	556	556	556	556	555
Panel B: Privately Agree					
Treat	-0.00965	-0.0114	-0.0105	-0.00122	0.0000101
	(0.0416)	(0.0415)	(0.0403)	(0.0382)	(0.0371)
Mean	0.407	0.407	0.407	0.407	0.407
SD	0.492	0.492	0.492	0.492	0.492
IDs	315	315	315	315	315
Obs	769	769	769	769	769
Topic FE	✓	✓	✓	✓	✓
Baseline guesses		✓	✓	✓	✓
Session Controls			✓	✓	✓
Ind Controls				✓	✓

Notes: The outcome variable $Express = 1$ indicates if individual i publicly expresses their views truthfully on topic j. Combined results on four topics (renaming schools affirmative action, death penalty and immunizations) for second movers are reported. Session controls include time of the session, week of the session, group size and moderator fixed effects. Individual controls include gender, race and ethnicity, year in school, major and self-reported political ideology. Standard errors are clustered at the Zoom session level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

a socially inappropriate view. Again, as assumed in our model, no significant effects are found for those who privately hold the socially appropriate view.

Furthermore, our results indicate that beliefs about public opinion continue to diverge across the Treatment and Control groups following the Zoom discussions. Participants in the Treatment group infer that the socially appropriate view is 15% (10.4 percentage points) less prevalent compared to the Control group's beliefs at endline (see Columns 4-6 in Table 2), relative to a 10% (6.7 percentage point) difference between the Treatment and Control group at midline. The difference in estimated treatment effects between midline and endline is significantly different at the 1% level (p-value = 0.004). Alternatively, a difference-in-differences approach, which we report in Table C.10, also shows that our information treatment has approximately a 61% larger effect on endline beliefs relative to midline beliefs (p-value = 0.013).

Taken together, our results are consistent with dynamic, self-reinforcing effects. Increasing attention to silence decreases the perceived popularity of the socially appropriate view at midline, which encourages participants to publicly express the socially inappropriate view in group discussions, which further decreases the perceived popularity of the socially appropriate view at endline. Over time, Control students increasingly overestimate the popularity of

appropriate views while Treatment students' beliefs become more accurate.

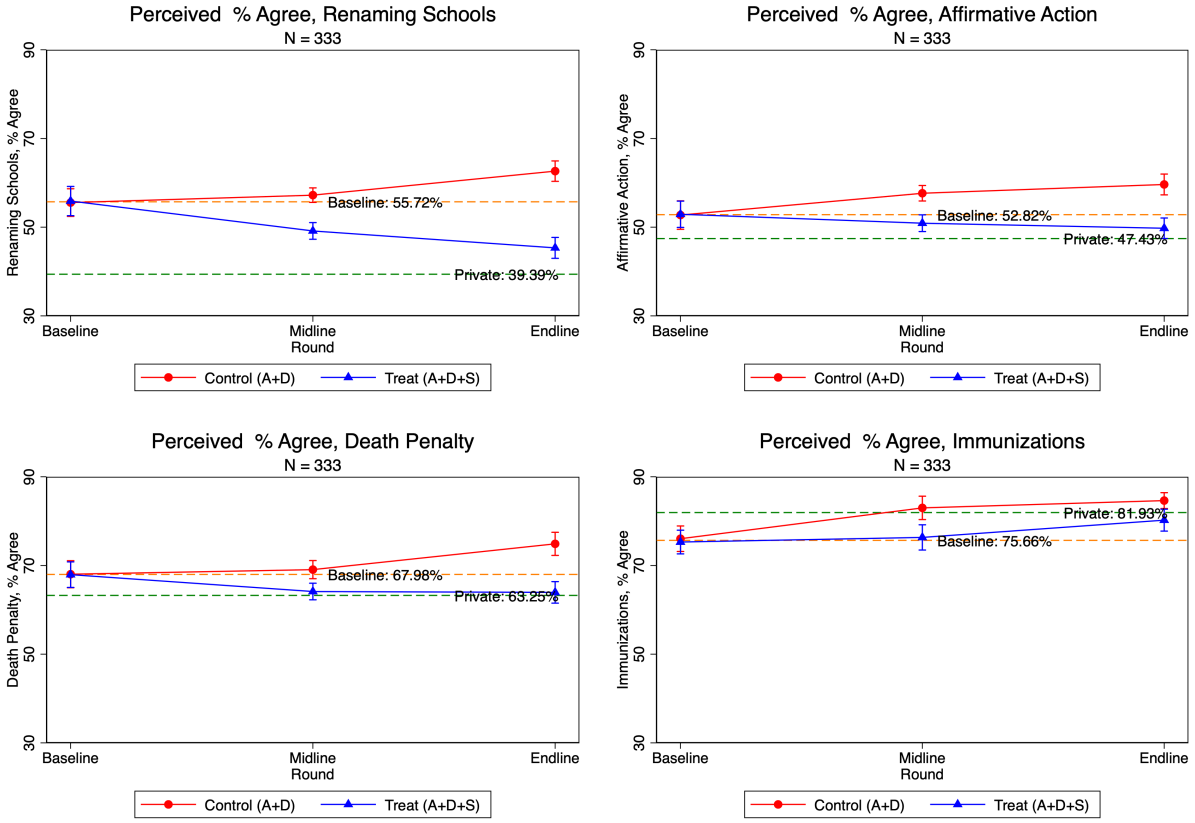


Figure 5: Dynamic Beliefs about %*Agree* Over Time

Note: The graphs show participants' average guesses about the share of agreement for each statement. We report beliefs elicited at baseline, midline (after the information treatment and before the Zoom session), and endline (after the Zoom session) respectively. The dashed green lines show the fraction of participants who privately agree with each statement at baseline. The dashed orange line shows the perceived %*Agree* among Control participants at baseline. 95% confidence intervals are included.

This dynamic belief-updating process is visualized in Figure 5, which illustrates the gradual divergence in perceived belief distributions between the Control and Treatment groups. For the three topics where participants initially overestimated the share of agreement in the baseline survey (Affirmative Action, Death Penalty, and Renaming Schools), the Treatment group progressively converges towards the true belief distribution. In contrast, the Control group diverges further from the true belief distribution overtime, increasing their overestimation of the popularity of the socially appropriate view on average. It appears that the Control information does not act as a neutral information treatment but rather moves the beliefs of Control participants towards the views publicly expressed by First Movers. As a result, endline beliefs among Control participants are significantly higher than their baseline beliefs (p-values $< .01$ for all four topics).

For the Immunizations topic, recall that participants actually underestimated the pop-

ularity of the socially acceptable view at baseline. However, we observe similar patterns in inference and expression, corresponding to the model predictions. First, individuals who hold the socially acceptable view are more likely to publicly express their views, although this difference is not statistically significant. Second, drawing attention to silence reduces the perceived popularity of the socially appropriate view for the Treatment group relative to the Control group at midline and at endline.

The combined results on expression and inference suggest that Control participants experience a spiral of silence. They form perceptions based only on expressed opinions, which leads them to overestimate the prevalence of the socially appropriate view. This overestimation further discourages them from expressing the socially inappropriate view, which reinforces the perception that the socially appropriate view is more popular than it actually is. Drawing attention to silence effectively breaks this spiral. The Treatment information causes participants to pay more attention to silence when observing the public discourse of the First Movers, which causes them to (correctly) adjust their beliefs downward about the level of popular support for the socially appropriate view, which increases their willingness to voice the socially inappropriate view publicly, which causes them to (correctly) adjust their beliefs further downward about the true popularity of the socially appropriate view. At endline, Treatment participants' beliefs about public opinion have become more accurate while Control participants' beliefs diverge further from the truth.

In Appendix D, we present additional analysis testing the robustness of our main results to different control variables, measuring spillover effects on other topics not discussed in the Zoom sessions, and allowing for beliefs to be multi-modal.

5 Mechanisms

In this section, we provide evidence for the mechanisms underlying our main experimental results, including the importance of social norms in generating self-censorship, identifying *attention* to silence as a key mechanism driving treatment effects, exploring whether our experimental results are consistent with alternative belief formation models, and the importance of omitting silence as opposed to other types of information.

5.1 Placebo Topic

To isolate social norms as a key driver of self-censorship and subsequent misperceptions, we use Daylight Saving Time as a placebo topic. This topic is largely socially neutral, as suggested by the Krupka–Weber norm-elicitation survey, in which "neutral" is the modal

response among Berkeley students. Our conceptual framework predicts that there should be no spiral of silence for topics that do not have socially sanctioned viewpoints. Corresponding to this prediction, we find no treatment effects for this placebo topic, neither on inference nor expression (Figures D.3 and D.4).

Formally, we estimate treatment effects using difference-in-differences with the placebo topic serving as a benchmark and report the results in Tables D.12 and D.13. The coefficient of Treat $\times$ Controversial Topic captures how the treatment affects beliefs and expression decisions on the controversial topics relative to the placebo benchmark. We find statistically significant differential effects across these two types of topics. For socially sensitive topics, the interaction term is negative at both midline and endline, indicating that directing participants' attention to silence reduces beliefs about the dominance of the socially appropriate view for Controversial topics relative to the placebo (Table D.12). For expression decisions (Table D.13), the interaction term is positive among participants who privately disagree with the socially appropriate view, suggesting that the treatment increases the willingness to publicly dissent on Controversial topics relative to the placebo.

These results should be viewed as suggestive given that only one placebo topic was discussed, making the placebo topic more vulnerable to idiosyncratic topic-specific features. Nevertheless, these patterns are consistent with social norms distorting expression and inference for socially sensitive topics but not for socially neutral topics

5.2 Recall and Interpretation of Silence

To examine whether *attention* to silence is a key mechanism underlying the observed treatment effects, we compare the accuracy of participants' recall of the Zoom discussions they participated in. In our endline survey, we ask students questions, incentivized for accuracy, about how many participants attended their Zoom session, how many participants expressed views agreeing or disagreeing with each discussion topic, and how many participants stayed silent during each discussion topic. We interpret the accuracy of recall as a proxy measure for the level of attention paid to these various aspects of the Zoom discussions.[16]

Our results indicate that Treatment students have more accurate recall of the number of participants who stayed silent during each Zoom discussion topic. On average, 63.7% of

[16]Note that we do not equate correct recall about the number of silent participants as full attention to silence due to the structure of the survey. We asked questions about recall after we elicited endline beliefs about the views of others participants. Mechanically, if participants noticed the total number of Zoom participants and could correctly recall the number of students who expressed views agreeing and disagreeing with each topic statement, they could calculate the number of participants who stayed silent even if they did not pay attention to silent participants during the Zoom session. Thus, we consider these recall questions as proxy measures rather than actual measures of attention.

Treatment participants correctly recalled the number of participants who stayed silent during the Zoom discussion, compared to 49.3% of Control participants (p-value < 0.001).[17]

How participants update their beliefs from silence depends not only on the level of attention they pay to silence but also on how they interpret that silence. If participants correctly infer the direction of selection bias into silence, then our model predicts that a treatment that increases attention to silence should significantly change subsequent inference and expression.

To test for this sophistication about selection bias into silence, in our endline survey we also ask participants to guess the private views of individuals who stayed silent on each topic. Among those who remain silent in Zoom sessions, on average 51% privately agree with the socially appropriate view, which is significantly lower than the expressed share of agreement (65%). Table C.11 reports the distribution of private beliefs by silence versus expression for each topic. We classify a respondent as "sophisticated" if they believe that the share of disagreement is higher among the silent subsample than among the expressed subsample. Our results suggest that the treatment effects on endline beliefs are approximately 3.2 percentage points (40%) stronger for those who understand that silence is selected, consistent with such sophistication being another key mechanism driving observed treatment effects (see Column 2 in Table D.14). These results are robust to alternative definitions of sophistication, such as classifying students as sophisticated if they correctly guess the empirical share of silent participants who privately disagree for each topic, rounded to the nearest 10th percentile.[18]

5.3 Alternative Interpretations of Belief Updating

In our conceptual framework, we assumed Bayesian updating for tractability, but in reality individuals may update from new information in varied ways. Salience-based belief formation models such as Bordalo et al. (2025) suggest that, while there are many different pieces of information available, individuals may pay attention to those that are salient and ignore all others when forming their beliefs. In fact, the key parameter in our model, attention to silence (λ), can be interpreted either in a Bayesian way or in a salience-based way. Under a Bayesian interpretation, λ captures the weight placed on specific pieces of information. Under a salience-based interpretation, individuals either do or do not attend to silence ($\lambda \in \{0, 1\}$),

[17] We consider an answer to be approximately correct if it falls within the $[-1, 1]$ window of the correct number of silent participants.

[18] Another notable pattern is that students who privately disagree with the socially appropriate view are more likely to infer that those who stay silent hold the socially inappropriate view. Holding all else constant, those who privately disagree with the socially appropriate views are 9 percentage point more likely to be sophisticated about the selection bias into silence. Hence treatment effects are also stronger for those who privately disagree with the socially appropriate view (Table D.14). Table D.14 also reports heterogeneous treatment effects for other demographics, including gender, race, ideology and major. The treatment effects are in general weaker for Asian students.

and the attention level is triggered by salience. Our treatment makes silence salient, thus making it more likely for participants to attend to this feature. Under this interpretation, the attention parameter can capture salience rather than rational information weighting, and the treatment operates by shifting which features of the environment individuals attend to.

We therefore conduct supplementary analysis to explore whether our treatment operates by increasing attention to silence, rather than by uniformly shifting beliefs. To do so, we examine whether belief distributions exhibit latent heterogeneity consistent with a mixture of silence-attentive and silence-inattentive types. We formalize this intuition using two-component finite mixture models, which fit the data better than a single normal and identify two latent belief types that differ in their guesses about the share of agreement in the midline survey. Across topics, the treatment significantly increases the probability of belonging to the lower-belief type, which we interpret as the silence-attentive type, suggesting that the intervention primarily affects the composition of belief types. To validate this interpretation, we link this latent-type classification to participants' recall of silence in the Zoom discussions. We find that participants classified as silence-attentive are significantly more likely to correctly recall silence. These findings suggest that the information treatment operates by increasing the share of silence-attentive individuals, leading to lower perceptions about the share of agreement and corresponding changes in expression decisions. A detailed description of this analysis is provided in Appendix D.3. Thus, while we use Bayesian updating as a tractable benchmark throughout, our model predictions and experimental results are consistent with both Bayesian updating and with salience-based models of belief formation.

5.4 Polarization

Thus far, we have assumed that opinions are binary. However, opinions are more continuous in reality. Indeed, Likert scales ranging from "strongly agree" to "strongly disagree" are often used in public opinion polls. We therefore conduct an additional experiment (N = 158) where we allow private and public expression for each topic to occur on a 5-point scale, ranging from strongly disagree to strongly agree, to allow participants to convey not only the direction but also the strength of their opinions. We follow a similar design to our main experiment, except that public expression now takes the form of survey responses that are publicly shared with other study participants. Details of this additional experiment are included in Appendix E.2.

By allowing opinions to take on a wider range of values, we find that silence is now selected by both the social appropriateness of views and the extremity of views (Table E.17). Public expression is skewed in the direction of socially appropriate views and is more heavy-tailed

than the actual belief distribution. Next, we find that exogenously drawing attention to silence reduces beliefs about the prevalence of both socially appropriate views and extreme views, with only marginal significance for extreme views (Table E.18). This in turn makes Treatment students more willing to express disagreement and more moderate views publicly compared to Control students, reducing observed polarization (Table E.19).

The results from this extension show that silence can be selected along multiple dimensions and that increasing attention to silence can correspondingly correct misperceptions and change public expression in multi-faceted ways.

5.5 Omitting Other Types of Information

In another extension, we use real data from a public feedback process conducted by UC Berkeley's Building Name Review Committee to replicate our main experimental results. We also introduce a third treatment arm to test whether silence is a larger source of misperceptions compared to other types of information omission. This additional experiment is described in detail in Appendix E.2.

The Committee evaluates proposals submitted by university community members to rename existing campus buildings, and each proposal undergoes a period of open public feedback. Community members can share their comments either publicly on the Committee's website, or keep their comments confidential and visible only to the Committee. For example, during the most recent proposal to rename Moses Hall,[19] the committee collected 154 responses in total through their website's feedback form, 84 of which were public and 70 of which were confidential.[20]

Corresponding to our main experiment, we share summaries of this public discussion with students and ask them to guess the true belief distribution and whether they would be willing to sign a public petition corresponding to their private beliefs. A key difference is that we randomly assign students (N = 255) to one of three groups: Control, which omits the number of confidential comments; 2) Treatment, which includes the number of all three types of comments; or 3) Omit Agree, which omits the number of public agreement comments. As in the main experiment, we inform students in all three groups of the total number of comments. Thus, the number of comments belonging to the omitted category can be inferred, at least in theory, and the total amount of information available across groups is the same. The only difference across groups is which expression categories are more salient in the summary statistics shared with participants (Figure E.9).

[19] https://chancellor.berkeley.edu/task-forces/building-name-review-committee/building-name-review-moses-hall/building-name-review

[20] Note that no information is available regarding the content or direction of the 70 confidential comments.

Comparing the Control group to the Treatment group allows us to test if our main experimental results replicate in a setting where we use real data from a real public feedback process and measure outcomes on a different type of expression decision, namely willingness to sign a petition. Comparing Omit Agree to the Control and Treatment groups allows us to test whether inattention to silence, as opposed to information omission more generally, is a particularly important source of misperceptions and self-censorship.

Our results first replicate the findings from our main experiment (Table E.20). Participants in the Treatment group estimate on average that 12% (8.1 percentage points) fewer students privately agree with renaming relative to the Control group. Their guesses are thus closer to the actual share of agreement, which is 48%. Furthermore, those who privately disagree with renaming in the Treatment group are over four times more likely (a 20 percentage point increase) to sign a petition against renaming, relative to their counterparts in the Control group. These results suggest that increasing attention to silence not only affects political discourse but also political action, such as the willingness to sign a petition.

Second, results from the additional treatment arm show that omitting silence has distinct effects on both inference and expression, compared to other types of information omission. Participants in the Omit Agree group make similar inferences about the true belief distribution as those in the Treatment group (Table E.20), which are 10.2% (6.9 percentage points) lower than the guesses of the Control group. Furthermore, participants who privately disagree in the Omit Agree group are more likely to sign a petition against renaming compared to their counterparts in the Control group, although these results are weakly significant. Overall, omitting the number of agreeing comments generates similar inference and expression decisions compared to when all information is included, suggesting that information omission per se does not automatically generate misperceptions. In contrast, omitting silence produces self-reinforcing effects on misperceptions and distorts public expression, as in our main experiment.

Taken together, these results highlight that silence, unlike explicit expressions of agreement or disagreement, is a less salient form of expression. When explicit agreement is excluded, individuals recognize the irregularity and actively pay attention to the omission, whereas the omission of silence is more easily ignored.

6 Robustness and Ecological Validity

6.1 Replication

As a robustness check, we conduct a replication of the experiment to test the sensitivity of our results to specific design choices of the information treatments (N = 261).[21] In particular, we modify the information given to both Control and Treatment participants so that the instructions clearly state that first movers could agree, disagree, or choose to stay silent. We also remove all data visualizations and instead simply state the summary statistics. These variations make silence more salient to participants in the Control group, generating a weaker information treatment that is potentially closer to realistic settings in which silence is noticeable but not highlighted. The revised information treatments are displayed in Figure E.10. Aside from the design of the information treatments, the structure and protocols of the rest of the replication were identical to the original experiment. One exception is that we updated one of the topics discussed in the Zoom sessions, swapping the original topic around Covid vaccines, which was not as relevant in 2025, for a new topic around legalizing marijuana.

The results from this replication show that our original results do not seem to be driven by the particular way that information is presented to students. Even after making these modifications to the information treatments, our main experimental results still hold: students in the treatment group are 1) more likely to express inappropriate views publicly, and 2) form more accurate beliefs about the true private opinions of their peers, relative to students in the control group.

Specifically, Tables E.21 and E.22 and Figure E.11 report the main results on inference and public expression decisions respectively. The results are consistent with those from the main experiment and similar in magnitude. Immediately after receiving the information intervention, Treatment students guess that the socially appropriate view is on average 9 percentage points (14%) less prevalent compared to Control students. This, in turn, makes Treatment students more willing to express socially inappropriate views publicly. Participants in the Treatment group who privately hold socially inappropriate views are about twice as likely (18 percentage point increase) to express their opinions relative to their counterparts in the Control group. As before, there are no significant treatment effects on expression for students who hold socially appropriate views. After participating in these discussions, the

[21] For this replication, we had targeted a similar sample size as in our original experiment. However, we faced some challenges with recruiting study participants and were only able to recruit 353 participants, of which 261 completed all components of the study. Anecdotally, students seemed more reluctant to participate in a study on public expression given changes in the political climate between the original experiment, which took place in spring 2023, and the replication, which took place in fall 2025.

difference in beliefs between Treatment and Control participants further diverges. In the endline survey, Treatment students estimate that socially appropriate views are 14 percentage points (22%) less prevalent than Control student estimates.

6.2 Ecological Validity: ANES

We investigate the role of silence in political discourse and the broader opinion climate using representative national surveys on public opinion from the American National Election Studies (ANES). Specifically, we use ANES data to test the correlation between silence and misperceptions. Under the assumptions of our conceptual framework, the fraction of people who stay silent and the magnitude of misperceptions about the views of others should be positively correlated.[22] Intuitively, overestimation of the popularity of socially appropriate views could suppress expression of inappropriate views, and silence in turn can lead to less accurate beliefs about others.

Since the 1970s, the ANES elicits respondents' own beliefs about a series of political and socioeconomic issues, as well as their beliefs about Democratic and Republican parties' views on these issues. The survey asks respondents to place themselves, the Democratic party and the Republican party on a 7-point scale, for the following topics: *Left to right; Liberal to conservative; Government services-spending; Defense spending; Government-private medical insurance, Guaranteed job income, Aid to Blacks.* Taking the liberal-conservative scale as an example, the survey asks "We hear a lot of talk these days about liberals and conservatives. Here is a seven-point scale on the political views that people might hold, which are arranged from extremely liberal to extremely conservative: 1) Where would you place yourself on this scale? 2) Where would you place the Democratic Party on this scale? and 3) Where would you place the Republican Party on this scale?" Detailed survey questions can be found in Appendix E.4.

These questions allow us to construct misperception measures by comparing the actual ratings by Democratic survey respondents with the perceived ratings and similarly for Republicans. We construct two measures of misperceptions. The raw measure simply takes the absolute differences in the actual and perceived ratings, while the adjusted measure standardizes the direction so that positive misperceptions are consistent with party stereotypes. For example, the adjusted measure is positive if Democrats are thought to be more liberal than they actually are and if Republicans are thought to be more conservative than they actually are. The adjusted measure is our preferred measure because it considers the direction of misperceptions.

[22]The proposition on the positive correlation between silence and misperceptions is in Appendix A.3

To measure non-response rates, we code the dummy variable *non-response* as 1 if participants choose any of the following options when placing themselves on the 7-point scale: *Refused; Don't know; or Haven't thought much about this.* Despite the potential social desirability bias respondents might have when answering face-to-face and telephone surveys, ANES is a private survey. Therefore, we interpret the non-response rate as a lower bound for silence in public expression.[23]

Throughout our analysis, we also control for issue salience, as defined in Bordalo, Tabellini, and Yang (2021). Specifically, ANES has a question asking "What do you think are the most important problems facing the country?" We mapped the responses to this question to the topics covered (*Liberal or conservative; Government services-spending; Defense spending; Government-private medical insurance, Guaranteed job income, Aid to Blacks*) to code the topic-year-level salience measure.

Overall, we observe a positive correlation between the share of non-responses and the magnitude of misperceptions across the seven topics over the period of 1974-2020. As shown in Table E.23, after controlling for year fixed effects and topic fixed effects, as well as topic salience and actual beliefs, a 1% increase in the non-response rate is correlated with a 5% (0.2 percentage point) increase in misperceptions.

We do not interpret these coefficients as causal as both silence and misperceptions are endogenous and the causal link could work in both directions. For example, people could be more likely to stay silent when they misperceive social norms, and conversely, when more people stay silent, people could have less accurate beliefs about others. In fact, our main experiment indicates the plausibility of both scenarios. Instead of discussing the causal relationship, we highlight this positive correlation between silence and misperceptions. This is not a trivial correlation: it would be absent if we did not account for silence in political discourse, if silence was not influenced by private beliefs, or if individuals were fully attentive of silence. In other words, this correlation holds under a specific set of assumptions under which the spiral of silence forms. We view this as suggestive evidence in support of the ecological validity of our experimental results.

7 Conclusion

In this paper, we study how self-censorship and limited attention to silence interact to shape public discourse and social norms. Given recent discussions about self-censorship on

[23]The ANES started in 1948 and was conducted every two years before 2004 and every four years subsequently. Using varied methods of surveys, including face-to-face, telephone, and more recently, online, the ANES potentially incorporates a social desirability bias among respondents, especially those participating in direct interactions.

college campuses, we study this issue using experiments with UC Berkeley undergraduates. Our experiments reveal that perceived social norms shape public discourse on campus. Students who hold views they perceive to be socially inappropriate are more likely to stay silent in public discussion. After observing the public expression of other students, which is skewed towards the socially appropriate view, students increase their beliefs about the popularity of the socially appropriate view, become even more discouraged from expressing socially inappropriate views, and so on. In equilibrium, a spiral of silence occurs which enables misperceptions about social norms to persist indefinitely.

Our experiment and model show that inattention to silence is a key mechanism that perpetuates such a spiral and enables misperceptions to persist in equilibrium. When we experimentally increase attention to silence, students form more accurate beliefs about the true level of dissent, become more likely to voice opposing views to other students, which then leads to more accurate beliefs about public opinion. With sufficient attention to silence, the spiral of silence is broken and misperceptions are corrected.

Our study adds to a literature across the social sciences studying the powerful social forces that support conformity and persistence of the status quo. In particular, we identify a novel mechanism, inattention to silence, that can explain phenomena where social norms diverge from majority private preferences.

In doing so, our study suggests potential interventions that could foster more inclusive and informative public discourse on socially sensitive issues. In public forums, highlighting silence as an active type of expression can help observers form more accurate beliefs about true public opinion. For example, social media platforms could display the number of views next to the number of likes/dislikes and comments for each post. Likewise, when summarizing results from public discourse forums such as town hall meetings or opinion polls, including information about the number of participants who stayed silent or declined to answer a question can lead to more accurate perceptions about the true opinion climate.

To the extent that our beliefs about the views of others shapes our own actions, the accuracy of those beliefs has important implications for a variety of political and social outcomes. While our study focused on UC Berkeley undergraduate students, our findings may apply to any context where there are strong norms around socially sensitive issues. Future research could test whether increasing attention to silence affects inference and expression in more diverse empirical settings.

References

Barron, Kai, Steffen Huck, and Philippe Jehiel. 2024. "Everyday Econometricians: Selection Neglect and Overoptimism when Learning from Others." *American Economic Journal: Microeconomics* 16:162–98.

Benabou, Roland and Jean Tirole. 2011. "Laws and Norms." Working Paper 17579, National Bureau of Economic Research.

Bicchieri, Cristina. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *Quarterly Journal of Economics* 131 (4):1753–1794.

Bordalo, Pedro, John Conlon, Nicola Gennaioli, Spencer Kwon, and Andrei Shleifer. 2025. "How People Use Statistics." *The Review of Economic Studies* rdaf022.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice Under Risk." *Quarterly Journal of Economics* 127 (3):1243–1285.

———. 2013. "Salience and Consumer Choice." *Journal of Political Economy* 121 (5):803–843.

———. 2022. "Salience." *Annual Review of Economics, 2022* 14:521–544.

Bordalo, Pedro, Marco Tabellini, and David Yang. 2021. "Issue Salience and Political Steroetypes." .

Braghieri, Luca. 2024. "Political correctness, social image, and information transmission." *American Economic Review* 114 (12):3877–3904.

Brown, Jennifer, Tanjim Hossain, and John Morgan. 2010. "Shrouded Attributes and Information Suppression: Evidence from the Field*." *The Quarterly Journal of Economics* 125 (2):859–876.

Brundage, Matt, Andrew T Little, and Soosun You. 2024. "Selection Neglect and Political Beliefs." *Annual Review of Political Science* 27:63–85.

Burdea, Valeria and Jonathan Woon. 2022. "Online belief elicitation methods." *Journal of Economic Psychology* 90:102496.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2020. "From Extreme to Mainstream: The Erosion of Social Norms." *American Economic Review* 110 (11):3522–48.

Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais. 2017. "'Acting Wife': Marriage Market Incentives and Labor Market Investments." *American Economic Review* 107 (11):3288–3319.

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott. 2020. "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia." *American Economic Review* 110 (10):2997–3029.

Bursztyn, Leonardo and David Y Yang. 2021. "Misperceptions about Others." Working Paper 29168, National Bureau of Economic Research.

Bénabou, Roland and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5):1652–1678.

———. 2016. "Mindful Economics: The Production, Consumption, and Value of Beliefs." *Journal of Economic Perspectives* 30 (3):141–64.

C-SPAN, video. 1969. "President Nixon's "Silent Majority" Speech on Vietnam War." URL `https://www.c-span.org/video/?153819-1/president-nixons-silent-majority-speech-vietnam-war`.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review* 99 (4):1145–77.

College Pulse, FIRE RealClear. 2021. "2021 College Free Speech Rankings." URL `https://reports.collegepulse.com/college-free-speech-rankings-2021`.

Coudray, Sylvie. 2015. *UNESCO: freedom of expression, information and the media*. Cambridge University Press, 208–234.

Crosetto, Paolo, Antonio Filippin, Peter Katuščák, and John Smith. 2020. "Central tendency bias in belief elicitation." *Journal of Economic Psychology* 78:102273.

Danz, David, Lise Vesterlund, and Alistair J. Wilson. 2022. "Belief Elicitation and Behavioral Incentive Compatibility." *American Economic Review* 112 (9):2851–83.

Deversi, Marvin, Alessandro Ispano, and Peter Schwardmann. 2021. "Spin Doctors: An Experiment on Vague Disclosure." *European Economic Review* 139:103872.

Ekins, Emily E. 2020. "Poll: 62% of Americans Say They Have Political Views They're Afraid to Share." *Cato Institute, Survey Reports* .

Enke, Benjamin. 2020. "What You See Is All There Is." *The Quarterly Journal of Economics* 135 (3):1363–1398.

Esponda, Ignacio and Emanuel Vespa. 2018. "Endogenous sample selection: A laboratory study." *Quantitative Economics* 9 (1):183–216.

Fang, Ximeng, Lorenz Goette, Bettina Rockenbach, Matthias Sutter, Verena Tiefenbeck, Samuel Schoeb, and Thorsten Staake. 2020. "Complementarities in Behavioral Interventions: Evidence From a Field Experiment on Energy Conservation." Tech. rep., University of Bonn and University of Mannheim, Germany.

Fernández-Duque, Mauricio. 2022. "The probability of pluralistic ignorance." *Journal of Economic Theory* 202 (C).

Giglio, Stefano and Kelly Shue. 2014. "No News Is News: Do Markets Underreact to Nothing?" *The Review of Financial Studies* 27 (12):3389–3440.

Glynn, Carroll J., Andrew F. Hayes, and James Shanahan. 1997. "Perceived Support for One's Opinions and Willingness to Speak Out: A Meta-Analysis of Survey Studies on the "Spiral of Silence"." *The Public Opinion Quarterly* 61 (3):452–463.

Gonzenbach, William J. 1992. "The Conformity Hypothesis: Empirical Considerations for the Spiral of Silence's First Link." *Journalism Quarterly* 69 (3):633–645.

Graeber, Thomas. 2022. "Inattentive Inference." *Journal of the European Economic Association* 21:560–592.

Hager, Anselm, Elnura Kazakbaeva, Lukas Hensel, and Damir Esenaliev. 2025. "Mutual Knowledge of Social Norms and Political Behavior." Working paper.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. 2014. " Learning Through Noticing: Theory and Evidence from a Field Experiment." *The Quarterly Journal of Economics* 129 (3):1311–1353.

Hayes, Andrew F., Carroll J. Glynn, and James Shanahan. 2005. "Validating the Willingness to Self-Censor Scale: Individual Differences in the Effect of the Climate of Opinion on Opinion Expression." *International Journal of Public Opinion Research* 17 (4):443–455.

Hirshleifer, David and Siew Hong Teoh. 2003. "Limited attention, information disclosure, and financial reporting." *Journal of Accounting and Economics* 36 (1-3):337–386.

Jehiel, Philippe. 2018. "Investment Strategy and Selection Bias: An Equilibrium Perspective on Overoptimism." *American Economic Review* 108:1582–97.

Jin, Ginger Zhe, Michael Luca, and Daniel Martin. 2021. "Is No News (Perceived As) Bad News? An Experimental Investigation of Information Disclosure." *American Economic Journal: Microeconomics* 13 (2):141–73.

Karlan, Dean, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman. 2016. "Getting to the Top of Mind: How Reminders Increase Saving." *Management Science* 62:3393–3411.

Koehler, Jonathan J. and Molly Mercer. 2009. "Selection Neglect in Mutual Fund Advertisements." *Management Science* 55 (7):1107–1121.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213–236.

Krosnick, Jon A. and Stanley Presser. 2010. *Question and Questionnaire Design*. Emerald Group Publishing, 2 ed., 263–313.

Krupka, Erin L. and Roberto A. Weber. 2013. "Identifying Social Norms Using Coordination Games: Why does Dictator Game Sharing Vary?" *Journal of the European Economic Association* 11 (3):495–524.

Kuran, Timur. 1995. *Private Truths, Public Lies*. Harvard University Press.

Li, Xinxin and Lorin Hitt. 2008. "Self Selection and Information Role of Online Product Reviews." *Information Systems Research* 19:456–474.

Lifchits, George, Ashton Anderson, Daniel G Goldstein, Jake M Hofman, and Duncan J Watts. 2021. "Success Stories Cause False Beliefs about Success." *Judgment and Decision Making* 16:1439–1463.

Lopéz-Pérez, Raúl, Ágnes Pintér, and Rocío Sánchez-Mangas. 2022. "Some Conditions (Not) Affecting Selection Neglect: Evidence from the Lab." *Journal of Economic Behavior Organization* 195:140–157.

Matthes, Jörg. 2014. "Observing the "Spiral" in the Spiral of Silence." *International Journal of Public Opinion Research* 27 (2):155–176.

Matthes, Jörg, Johannes Knoll, and Christian von Sikorski. 2018. "The "Spiral of Silence" Revisited: A Meta-Analysis on the Relationship Between Perceptions of Opinion Support and Political Opinion Expression." *Communication Research* 45:3–33.

Montero, Maria and Jesal Sheth. 2021. "Naivety about Hidden Information: An Experimental Investigation." *Journal of Economic Behavior Organization* 192:92–116.

Moreno-Riano, Gerson. 2002. "Experimental implications for the Spiral of Silence." *Social Science Journal* 39 (1):65+.

Morris, Stephen. 2001. "Political Correctness." *Journal of Political Economy* 109 (2):231–265.

New York Times, a. 2022. "I Came to College Eager to Debate. I Found Self-Censorship Instead." URL `https://www.nytimes.com/2022/03/07/opinion/campus-speech-cancel-culture.html`.

Nickerson, Raymond. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2:175–220.

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

Pew Research Center, a. 2019. "Sizing up Twitter Users." URL `https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/`.

Przeworski, Adam, Susan Stokes, and Bernard Manin, editors. 1999. *Democracy, Accountability, and Representation*. Cambridge University Press.

Scheufle, Dietram A. and Patricia Moy. 2000. "Twenty-five Years of the Spiral of Silence: A Conceptual Review and Empirical Outlook." *International Journal of Public Opinion Research* 12:3–28.

Shamir, Jacob and Michal Shamir. 2000. *The anatomy of public opinion*. University of Michigan Press.

Spranca, Mark, Elisa Minsk, and Jonathan Baron. 1991. "Omission and Commission in Judgment and Choice." *Journal of Experimental Social Psychology* 27:76–105.

Stango, Victor and Jonathan Zinman. 2014. "Limited and Varying Consumer Attention: Evidence from Shocks to the Salience of Bank Overdraft Fees." *The Review of Financial Studies* 27 (4):990–1030.

Taubinsky, Dmitry and Alex Rees-Jones. 2017. "Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment." *The Review of Economic Studies* 85 (4):2462–2496.

Tversky, Amos and Daniel Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185:1124–31.

WBUR, Here & Now. 2020. "Made Famous By Nixon, The Phrase 'Silent Majority' Resurfaces For Trump's 2020 Reelection." URL `https://www.wbur.org/hereandnow/2020/07/23/what-is-silent-majority-trump-nixon`.

Yun, Gi Woong and Sung-Yeon Park. 2011. "Selective Posting: Willingness to post a message online." *Journal of Computer-Mediated Communication* 16:201–227.

Appendices

Appendix A Model: Proofs and Extensions

A.1 Proof of Proposition 1

Proof of Proposition 1. Consider:

$$\begin{aligned}\rho_j \;&=\; \frac{\mathbb{E}[\pi_{j1}|\mathcal{S}]\,G(0)}{\mathbb{E}[\pi_{j1}|\mathcal{S}]\,G(0) \;+\; (1-\mathbb{E}[\pi_{j1}|\mathcal{S}])\,G\big(\chi\,\rho_j\mathbb{E}[\pi_{j1}|\mathcal{S}]\big)} \\ &< \frac{\mathbb{E}[\pi_{j1}|\mathcal{S}]\,G(0)}{\mathbb{E}[\pi_{j1}|\mathcal{S}]\,G(0) \;+\; (1-\mathbb{E}[\pi_{j1}|\mathcal{S}])\,G(0)} = \mathbb{E}[\pi_{j1}|\mathcal{S}]\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\pi_{j1}|\mathcal{S}] = \frac{a+S_A+\rho_j S_S}{a+d+S_A+S_D+S_S} &< \frac{a+S_A+\mathbb{E}[\pi_{j1}|\mathcal{S}]S_S}{a+d+S_A+S_D+S_S} \\ \implies \mathbb{E}[\pi_{j1}|\mathcal{S}] &< \frac{a+S_A}{a+d+S_A+S_D} = \mathbb{E}[\pi_{j0}|\mathcal{S}]\end{aligned}$$

□

A.2 Proof of Proposition 3

Proof of Proposition 3. For simplicity, and with a slight abuse of notation, we omit the asterisks in the following proofs and use π_0 and π_1 as shorthand for the equilibrium beliefs π_0^* and π_1^*, respectively.

Consider the intrinsic value of expression V and the strength of social sanctions χ, we denote by $F(\cdot)$ the cumulative distribution function (CDF) of V/χ.

Regularity conditions

- Smooth inference: The CDF $F(\cdot)$ governing expression decisions is strictly increasing and continuously differentiable.
- Sufficient heterogeneity in intrinsic expression values: There exists a finite constant $L > 0$ such that $\sup_{t\in R} f(t) \leq L$ and L is sufficiently small.

Collect three equilibrium-defining equations:

$$\mathbb{E}[\pi_0] \;=\; \frac{p\,[\,1-F(0)\,]}{p\,[\,1-F(0)\,] + (1-p)\Big[m\big[1-F(\rho\,\mathbb{E}[\pi_1])\big] + (1-m)\big[1-F(\mathbb{E}[\pi_0])\big]\Big]}\,, \qquad (6)$$

$$\mathbb{E}[\pi_1] \;=\; p\,[\,1-F(0)\,] \;+\; \Big[p\,F(0)+(1-p)\,\big(m\,F(\rho\,\mathbb{E}[\pi_1])+(1-m)\,F(\mathbb{E}[\pi_0])\big)\Big]\,\rho, \qquad (7)$$

$$\rho \;=\; \frac{\mathbb{E}[\pi_1]\,F(0)}{\mathbb{E}[\pi_1]\,F(0)+(1-\mathbb{E}[\pi_1])\,F\big(\rho\,\mathbb{E}[\pi_1]\big)}. \qquad (8)$$

Equations (6)–(8) determine $\mathbb{E}[\pi_0]$ and $\mathbb{E}[\pi_1]$ as functions of $m \in [0,1]$, $p \in (0,1)$, and $F(\cdot)$.

Step 1. $\mathbb{E}[\pi_0] > \mathbb{E}[\pi_1]$ *for all* $m \in [0,1]$.

$\mathbb{E}[\pi_1] = \Pr(\theta_i = A \mid e_i = A)\ \Pr(e_i = A) + \Pr(\theta_i = A \mid e_i = 0)\ \Pr(e_i = 0) = (1-F(0))\,p + \Pr(e_i = 0)\,\rho.$

From (8), since $F(\rho\,\mathbb{E}[\pi_1]) > F(0)$ whenever $\rho > 0$, we get

$$\rho = \frac{\mathbb{E}[\pi_1]\,F(0)}{\mathbb{E}[\pi_1]\,F(0)+(1-\mathbb{E}[\pi_1])\,F(\rho\,\mathbb{E}[\pi_1])} < \frac{\mathbb{E}[\pi_1]\,F(0)}{\mathbb{E}[\pi_1]\,F(0)+(1-\mathbb{E}[\pi_1])\,F(0)} = \mathbb{E}[\pi_1].$$

Substitute $\rho < \mathbb{E}[\pi_1]$ into the expression for $\mathbb{E}[\pi_1]$:

$$\mathbb{E}[\pi_1] < (1-F(0))\,p + \Pr(e_i = 0)\,\mathbb{E}[\pi_1].$$

Rearranging,

$$\mathbb{E}[\pi_1]\big[1-\Pr(e_i=0)\big] < (1-F(0))\,p \implies \mathbb{E}[\pi_1] < \frac{1-F(0)}{1-\Pr(e_i=0)} = \frac{\Pr(e_i = A)}{\Pr(e_i \neq 0)} = \mathbb{E}[\pi_0].$$

Thus, for any m, individuals who ignore silence hold strictly higher equilibrium beliefs about the prevalence of socially appropriate views than individuals who pay attention to silence.

Step 2. $\mathbb{E}[\pi] = \mathbb{E}[\pi_1] = p$ *when* $m = 1$.
When $m = 1$, obviously $\mathbb{E}[\pi] = \mathbb{E}[\pi_1]$. Substitute $m = 1$ into (7):

$$\mathbb{E}[\pi_1] = p\,[\,1-F(0)\,] + \big[p\,F(0)+(1-p)\,F(\rho\,\mathbb{E}[\pi_1])\big]\,\rho.$$

Rearrange:

$$\mathbb{E}[\pi_1] - p = p\,(\rho-1)\,F(0) + (1-p)\,\rho\,F(\rho\,\mathbb{E}[\pi_1]),$$

i.e.

$$\big(\mathbb{E}[\pi_1]-p\big)\,\Big[1-\frac{F(0)\,F(\rho\,\mathbb{E}[\pi_1])}{\mathbb{E}[\pi_1]\,F(0)+(1-\mathbb{E}[\pi_1])\,F(\rho\,\mathbb{E}[\pi_1])}\Big] = 0.$$

Note

$$\frac{\mathbb{E}[\pi_1]}{F\big(\rho\,\mathbb{E}[\pi_1]\big)} + \frac{1-\mathbb{E}[\pi_1]}{F(0)} > \mathbb{E}[\pi_1] + 1 - \mathbb{E}[\pi_1] = 1$$

$$\Longrightarrow 1 - \frac{F(0)\,F\big(\rho\,\mathbb{E}[\pi_1]\big)}{\mathbb{E}[\pi_1]\,F(0) + [1-\mathbb{E}[\pi_1]]\,F\big(\rho\,\mathbb{E}[\pi_1]\big)} > 0,$$

Thus $\mathbb{E}[\pi_1] - p$ must be 0. Intuitively, when everyone pays full attention to silence, inference is guided by the same data generation rule which also determines the expression decisions, so the perceived belief distribution eventually converges to the actual belief distribution.

Step 3. $\mathbb{E}[\pi_0]$, $\mathbb{E}[\pi_1]$, $\mathbb{E}[\pi]$ *are all strictly decreasing in* $m \in [0,1]$.

Lemma A.2 (monotonicity of Silence Inference: $\rho(x)$).
For each $x \in (0,1)$, define $\rho(x)$ implicitly by

$$\rho(x) \;=\; \frac{x\,F(0)}{x\,F(0) + (1-x)\,F\big(\rho(x)\,x\big)}.$$

Under the regularity conditions above, $\dfrac{d\rho}{dx} > 0$.

Sketch of Lemma A.2. Define $\Phi(\rho, x) = \rho[xF(0)+(1-x)F(\rho x)] - xF(0) = 0$. By the implicit function theorem: $\rho'(x) = -\Phi_x/\Phi_\rho$.

$$\Phi_\rho = xF(0) + (1-x)F(\rho x) + \rho(1-x)xF'(\rho x) > 0$$

Therefore, it suffices to show $\Phi_x < 0$. We have

$$\begin{aligned}\Phi_x &= \rho(F(0) - F(\rho x) + (1-x)\rho F'(\rho x)) - F(0) \\ &= -\underbrace{[(1-\rho)F(0) + \rho F(\rho x)]}_{\geq 0} + \underbrace{\rho^2(1-x)F'(\rho x)}_{\geq 0}\end{aligned}$$

Hence $\Phi_x < 0$, provided the positive correction term is sufficiently small. A sufficient condition is a uniform bound $\sup_t f(t) \leq L$ with L small enough (together with interiority ensuring $\rho \leq \bar{\rho} < 1$), so that $\rho^2(1-x)F'(\rho x) \leq L < (1-\bar{\rho})F(0)$.

As a concrete example, if V/χ has a location–scale distribution $F(t) = \tilde{F}\big(\frac{t-\mu}{\sigma}\big)$ and $f(t) = F'(t) = \frac{1}{\sigma}\tilde{f}\big(\frac{t-\mu}{\sigma}\big)$ with bounded $\tilde{f}$, then $\sup_t f(t) = \frac{1}{\sigma}\sup_z \tilde{f}(z)$, so the condition $\sup_t f(t) \leq L$ holds whenever

$$\sigma \geq \bar{\sigma} \equiv \frac{\sup_z \tilde{f}(z)}{L}.$$

Intuitively, sufficient dispersion in intrinsic expression values dampens the responsiveness of expression behavior to beliefs, ensuring that changes in perceived agreement affect inference

from silence in a smooth and monotonic way. □

Step 3a. $\frac{d\,\mathbb{E}[\pi_1]}{dm}\,\big|_{m=1} < 0$.
Substitute ρ from (8) into (7) and rearrange:

$$\mathbb{E}[\pi_1] - p = \frac{(\mathbb{E}[\pi_1] - p)\,F(0)\,F(\rho\,\mathbb{E}[\pi_1]) + (1-p)\,\mathbb{E}[\pi_1]\,F(0)\,(1-m)\big[F(\mathbb{E}[\pi_0]) - F(\rho\,\mathbb{E}[\pi_1])\big]}{\mathbb{E}[\pi_1]\,F(0) + (1 - \mathbb{E}[\pi_1])\,F(\rho\,\mathbb{E}[\pi_1])}. \tag{9}$$

Define

$$H\big(\mathbb{E}[\pi_1],\,\mathbb{E}[\pi_0],\,\rho\big) \;=\; \frac{\mathbb{E}[\pi_1]\,F(0) + (1 - \mathbb{E}[\pi_1])\,F(\rho\,\mathbb{E}[\pi_1]) - F(0)\,F(\rho\,\mathbb{E}[\pi_1])}{(1-p)\,\mathbb{E}[\pi_1]\,F(0)\,\big[F(\mathbb{E}[\pi_0]) - F(\rho\,\mathbb{E}[\pi_1])\big]},$$

so (9) can be rewritten as

$$\big(\mathbb{E}[\pi_1] - p\big)\,H\big(\mathbb{E}[\pi_1],\,\mathbb{E}[\pi_0],\,\rho\big) \;=\; 1 - m.$$

Here $H(\cdot) > 0$. Differentiate both sides with respect to m:

$$(\mathbb{E}[\pi_1] - p)\,\frac{dH}{dm} + \frac{d\,\mathbb{E}[\pi_1]}{dm}\,H = -1.$$

At $m = 1$, by Step 2 we have $\mathbb{E}[\pi_1] = p$, so

$$\frac{d\,\mathbb{E}[\pi_1]}{dm}\Big|_{m=1} = -\frac{1}{H} < 0.$$

Thus $\mathbb{E}[\pi_1]$ is strictly decreasing at $m = 1$.

Step 3b. *Strict monotonicity of* $\mathbb{E}[\pi_1]$ *for all* $m \in [0, 1]$.

Rearrange (6) as

$$\Big(\frac{1}{\mathbb{E}[\pi_0]} - 1\Big)\,\frac{p\,(1 - F(0))}{1-p} \;=\; 1 - m\,F\big(\rho\mathbb{E}[\pi_1]\big) - (1-m)\,F\big(\mathbb{E}[\pi_0]\big).$$

Differentiate both sides with respect to m, treating $\mathbb{E}[\pi_0]$ and $\mathbb{E}[\pi_1]$ as functions of m. By Step 1,

$$F\big(\rho\,\mathbb{E}[\pi_1]\big) < F\big(\mathbb{E}[\pi_0]\big),$$

and since F is increasing and $\rho(\cdot)$ is increasing by Lemma A.2, $\partial_{\mathbb{E}[\pi_1]} F\big(\rho\,\mathbb{E}[\pi_1]\big) > 0$, we can

rewrite:

$$\underbrace{\Big(\frac{p\big(1-F(0)\big)}{\mathbb{E}[\pi_0]^2\,(1-p)} \;-\; (1-m)\,F'(\mathbb{E}[\pi_0])]\Big)}_{\text{denoted as } Q}\;\frac{\partial \mathbb{E}[\pi_0]}{\partial m}$$

$$= \underbrace{F\big(\rho\,\mathbb{E}[\pi_1]\big) \;-\; F\big(\mathbb{E}[\pi_0]\big)}_{<0} + m\;\underbrace{\frac{\partial F\big(\rho\,\mathbb{E}[\pi_1]\big)}{\partial \mathbb{E}[\pi_1]}}_{>0}\;\frac{\partial \mathbb{E}[\pi_1]}{\partial m}. \tag{10}$$

A uniform sufficient condition is:

$$\sup_t f(t) < \frac{p(1-F(0))}{1-p}$$

Again, this is the assumption that the there exists sufficient heterogeneity in the intrinsic value of expression. Intuitively, Bounding $\sup f$ says the distribution of expression values is not too concentrated anywhere.

To prove strict monotonicity of $\mathbb{E}[\pi_1]$ over $m \in [0,1]$, suppose, toward a contradiction, that it does not hold. Then, there exists some $\hat{m} \in [0,1]$ such that:

$$\frac{d\,\mathbb{E}[\pi_1]}{d\,m}\bigg|_{m=\hat{m}} \;=\; 0.$$

From equation (10), we then know

$$\frac{d\,E[\pi_0]}{d\,m}\bigg|_{m=\hat{m}} \;\neq\; 0.$$

Hence, there is a small neighborhood $(\hat{m}-\epsilon,\, \hat{m}+\epsilon)$ in which the function $E[\pi_1](m)$ is not monotonic, whereas $E[\pi_0](m)$ remains monotonic. Consequently, one can find two distinct mixing parameters

$$m,\, m' \;\in\; (\hat{m}-\epsilon,\, \hat{m}+\epsilon), \quad m \neq m',$$

such that

$$\mathbb{E}[\pi_1](m) \;=\; \mathbb{E}[\pi_1](m'), \quad \rho \;=\; \rho'.$$

Substituting these equalities into equation (7) yields

$$\big(F\big(\mathbb{E}[\pi_0](m)\big) \;-\; F\big(\rho\mathbb{E}[\pi_1](m)\big)\big)\,(1-m) \;=\; \big(F\big(\mathbb{E}[\pi_0](m')\big) \;-\; F\big(\rho'\,\mathbb{E}[\pi_1](m')\big)\big)\,(1-m').$$

Without loss of generality, assume $m < m'$. Then it must be that

$$\mathbb{E}[\pi_0](m) \;<\; \mathbb{E}[\pi_0](m').$$

Since $\frac{d\,\mathbb{E}[\pi_0]}{d\,m}\Big|_{m=\hat{m}} \neq 0$ on that interval, we have:

$$\frac{d\,\mathbb{E}[\pi_0](m)}{d\,m} \;>\; 0 \quad \text{for } m \in (\hat{m}-\epsilon,\, \hat{m}+\epsilon).$$

Choose $\bar{m} \in (\hat{m}-\epsilon,\, \hat{m}+\epsilon)$ such that

$$\frac{d\,\mathbb{E}[\pi_1]}{d\,m}\Big|_{m=\bar{m}} \;<\; 0, \quad \frac{d\,\mathbb{E}[\pi_0]}{d\,m}\Big|_{m=\bar{m}} \;>\; 0.$$

Substituting these into equation (10) produces a contradiction. Therefore,

$$\frac{d\,\mathbb{E}[\pi_1](m)}{d\,m} \;\neq\; 0, \quad \forall\, m \in [0,1].$$

Step 3c. $\mathbb{E}[\pi_1]$ *and* $\mathbb{E}[\pi_0]$ *are strictly decreasing in* m*.*

Combining Step 3a ($\mathbb{E}[\pi_1](m)$ is decreasing at $m=1$) and Step 3b (strict monotonicity of $\mathbb{E}(\pi_1)$ for all $m \in [0,1]$), we conclude

$$\frac{d\,\mathbb{E}[\pi_1](m)}{d\,m} \;<\; 0, \quad \forall\, m \in [0,1].$$

Plugging this into equation (10) yields

$$\frac{d\,\mathbb{E}[\pi_0](m)}{d\,m} \;<\; 0, \quad \forall\, m \in [0,1].$$

Step 3d. *Monotonicity of the overall belief* $\mathbb{E}[\pi]$*.*

Finally,

$$\mathbb{E}[\pi](m) \;=\; (1-m)\,\mathbb{E}[\pi_0](m) \;+\; m\,\mathbb{E}[\pi_1](m),$$

Differentiating and using Steps 1 and 3b-3c gives

$$\frac{d\,\mathbb{E}[\pi](m)}{d\,m} \;<\; 0, \quad \forall\, m \in [0,1],$$

completing the proof of Proposition 3. □ □

A.3 Proposition 4

The last prediction is a positive correlation between silence and misperceptions.

Proposition 4. *Assuming* $E[\pi] > p$ *and* $\lambda \in [0,1)$*, in period* t *when* $E[\pi_t] \neq E[\pi^*]$*, mis-*

perception ($\Delta_t \equiv E[\pi_t] - p$) is positively correlated with share of silence: $Corr(S_{S,t}, \Delta_t) > 0$.

Proof: Silence and Misperceptions in period t:

$$\begin{aligned} Cov(S_S, \Delta) &= Cov(pG(0) + (1-p)[mG(\chi\rho E[\pi]) + (1-m)G(\chi E[\pi]), E[\pi] - p) \\ &= m(1-p)Cov(G(\chi\rho E[\pi]), E[\pi]) + (1-m)(1-p)Cov(G(\chi E[\pi]), E[\pi]) > 0 \end{aligned}$$

Intuitively, overestimation about the share of agreement could stifle public expression, and silence leads to less accurate beliefs about others. Both silence and misperceptions are endogenous variables in our model, so instead of discussing the causal relationships between these two variables, we make cross-topic comparisons and test the positive correlation predicted by Proposition 4 with ANES data.

A.4 Other Ways of Updating

For tractability, we assumed Bayesian updating in the model. In this section, we examine other heuristic updating methods that individuals might employ. For example, individuals might apply a 50/50 rule that treats silence as 50% Agree, 50% Disagree; project their prior beliefs onto those who remain silent; or project their own private type onto silence. Figure A.1 displays the simulation results for these alternative updating rules. The simulation results highlight several key insights.

First and most importantly, across all updating rules we consider, the perceived popularity of the socially appropriate view exhibits a monotonic comparative static in the share of individuals who pay attention to silence. When a larger fraction of individuals pay attention to silence (higher m), the society as a whole infers a lower share of agreement $E[\pi]$. This monotonic relationship, which is the key prediction that we test experimentally, does not depend on a specific belief-updating rule. Second, when all individuals pay full attention to silence ($m = 1$), convergence toward the true belief distribution requires at least a partial understanding of the direction of self-selection in silence. Under the 50/50 updating rule, or under the simple projection of one's prior beliefs onto silence, beliefs fail to converge to the true belief distribution and instead systematically overestimate the prevalence of socially appropriate views.

Finally, in the experiment we found that individuals who privately disagree infer that those who remain silent are more likely to privately disagree. Motivated by this pattern, we simulate a "type projection" updating rule in which individuals map their private type (Agree or Disagree) onto silence. Belief trajectories under this rule closely resemble those obtained

under Bayesian updating. However, these predictions only hold under type projection if the true share of disagreement is greater than agreement (as in this simulation). However, experimentally, we find empirical support for these predictions for all four topics, regardless of whether the true share of disagreement is greater than or less than the share of agreement. Thus, type projection cannot explain all of the empirical results we find.

Taken together, these simulation exercises strengthen the interpretation of our experimental results and clarify the role of belief updating in the framework. Crucially, the key prediction that we test experimentally (the perceived $\%Agree$ decreases in the fraction of individuals who pay attention to silence m) does not depend on full Bayesian updating.

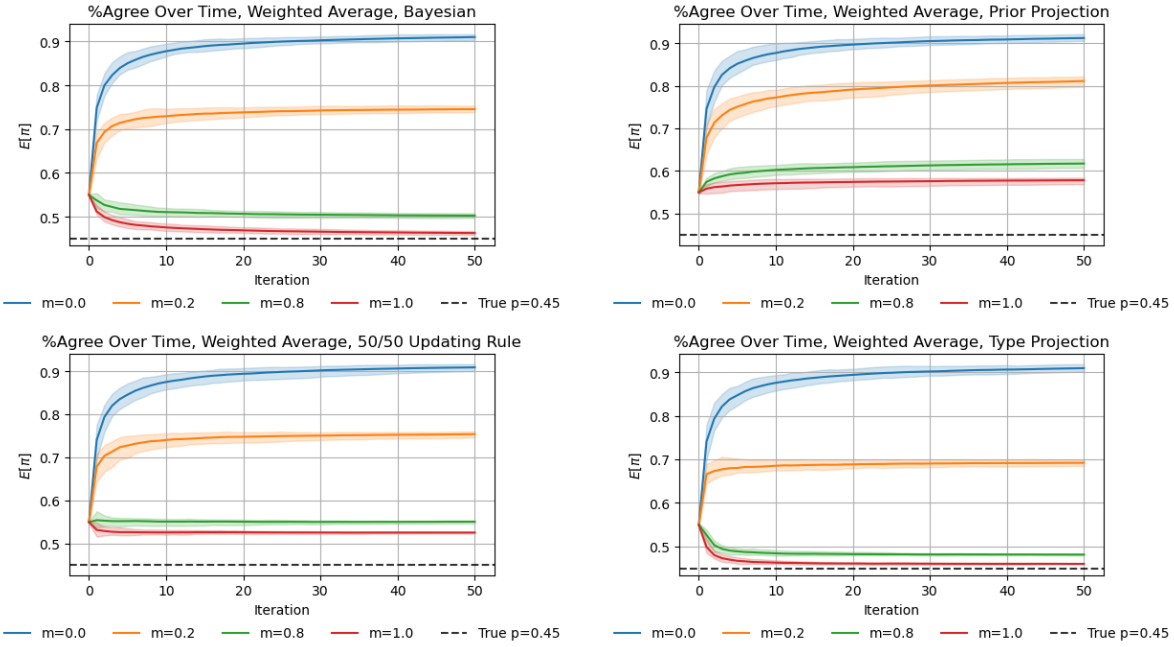


Figure A.1: Simulation of Beliefs Over Time Under Different Updating Rules

Note: This figure reports simulated belief dynamics for the perceived share of agreement, $\%Agree$, under four alternative belief-updating rules: Bayesian updating; a 50/50 rule that treats silence as 50% Agree and 50% Disagree; projection of one's prior beliefs onto those who remain silent; and projection of one's own private type onto silence. Each simulation consists of $T = 50$ iterations with a population of $N = 1000$ individuals, of whom 45% privately agree with the socially appropriate view and 55% privately disagree. m denotes the fraction of individuals who pay full attention to silence in the society. The horizontal dashed line denotes the true share of agreement in the population. The shaded areas indicate 95% simulation intervals.

Appendix B Discussion: Topic Selection

B.1 Topics Covered in the Baseline Survey

- Affirmative Action: If Proposition 209 was repealed, universities in the UC system should adopt extensive affirmative action policies that explicitly take into account race in the admission process.
- Criminalization: Non-violent crimes should be punishable with alternatives to jail.
- Death Penalty: The U.S. should abolish the death penalty.
- Daylight Saving Time "DST" (Placebo): Daylight saving time should be eliminated.
- Immigration: The U.S. should create more accessible pathways to citizenship, even for those who arrived illegally.
- Immunizations: Immunizations, such as for Covid and flu, should be required on Berkeley's campus.
- Job Requirements (Placebo): State government jobs should eliminate college degree requirements.
- Religious Values: Under the principles of freedom of religion, people have the right to make choices about who they serve or employ based on their religious values
- Renaming Schools: All public schools named after controversial historical figures, including former Presidents George Washington, Thomas Jefferson and Abraham Lincoln, should be renamed.
- Transgender Athletes: Transgender athletes should be allowed to compete in female sports at the college level.

B.2 Perceived Social Appropriateness of Topics

Table B.1: Perceived Social Appropriateness of each Statement

Topic	Very Inappropriate	Somewhat Inappropriate	Neutral	Somewhat Appropriate	Very Appropriate
Affirmative Action	6.32	23.16	21.05	**37.89**	11.58
Criminalization	11.58	25.26	24.21	**26.32**	12.63
Death Penalty	2.11	10.53	11.58	**45.26**	30.53
DST (P)	3.16	6.32	**43.16**	14.74	32.63
Immigration	3.16	5.26	9.47	**51.58**	30.53
Immunizations	12.63	12.63	9.47	**34.74**	30.53
Job Requirements (P)	5.26	18.95	30.53	**35.79**	9.47
Religious Values	14.74	10.53	7.37	31.58	**35.79**
Renaming Schools	2.11	16.84	17.89	**49.47**	13.68
Transgender Athletes	5.26	10.53	7.37	35.79	**41.05**

Notes: The table shows the percentage of Berkeley students who choose each respective level of social appropriateness for each statement. Students were incentivized to guess the modal responses. The actual modal response to each topic is indicated in bold.

B.3 Discussion

Figure B.2 shows the actual percentage of Berkeley student respondents who privately choose "agree" for each statement, along with students' guesses about the percentage of students who choose "agree". From these results, we observe that misperceptions are widespread: there exist statistically significant differences between actual and perceived belief distributions for 8 out of the 10 topics. Among these eight cases of misperceptions, four appear to be cases where participants significantly overestimate the fraction of people who hold the socially appropriate view, consistent with the spiral of silence theory. For the remaining four cases of significant misperceptions, participants appear to actually underestimate the fraction of people who hold the socially appropriate view, perhaps reflecting center-biased guessing (Crosetto et al., 2020; Danz, Vesterlund, and Wilson, 2022). Understanding when and how spiral of silence effects may interact with or dominate center-biased guessing or other behavioral biases is beyond the scope of this paper. However, these results do suggest that, while not universal, overestimation of the popularity of socially acceptable views is common and occurs somewhat frequently across a broad range of socially sensitive topics.

Due to time constraints and feasibility, we could only include a subset of these topics in the Zoom discussions. As our study is focused on testing the spiral of silence, we focus on topics where participants overestimate the popularity of the socially appropriate view. Thus, our study results should not be seen as representative of the universe of potentially socially sensitive topics, but rather is specific to topics where people misperceive the popularity of the socially appropriate view, and can be seen as upper bound estimates of the potential treatment effects of increasing attention to silence.

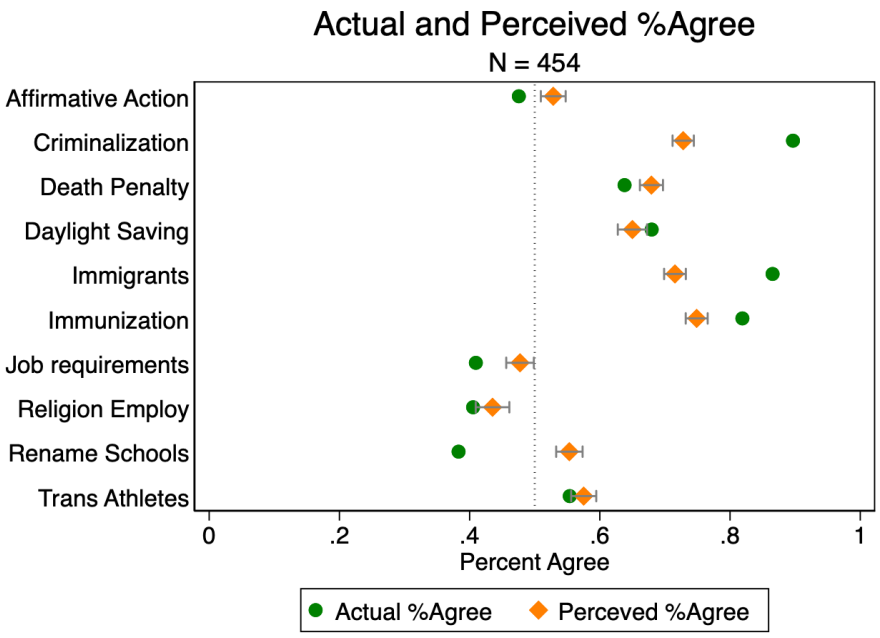


Figure B.2: Baseline Actual and Perceived Fraction of Student Agreement

Notes: In this figure, the green dot shows the actual share of agreement among all private views expressed by participants in the baseline survey. The orange diamond represents participants' guesses, incentivized for accuracy, of the share of agreement among all other participants for each of the topics in the baseline survey, with 95% confidence intervals displayed.

To satisfy these criteria, we chose the topics on Affirmative Action, Death Penalty, and Renaming Schools to be discussed in the Zoom sessions. For example, on average participants estimated that 56% of their fellow Berkeley student respondents would privately agree that all public schools named after controversial historical figures should be renamed, when in reality only 39% of respondents privately agreed with this statement. In a similar vein, for the Affirmative Action topic the perceived level of agreement was 53%, while the actual share of agreement was only 46%.

For the Zoom discussions, we also selected the Immunizations topic to represent a case where perceptions were center-biased, and we selected the Daylight Saving Time topic to serve as a placebo because it appears to be socially neutral. In total, the following five topics were discussed in all Zoom sections in the following order: Daylight Saving Time, Renaming Schools, Death Penalty, Affirmative Action, and Immunizations.

Appendix C Supplementary Figures and Tables

Table C.2: Summary Statistics

	Whole Sample (N = 383)	**First Movers** (N = 50)	**Control** (N = 166)	**Treatment** (N = 167)	**p Value** 1st/2nd Movers	**p Value** Control/Treat
Panel A: Demographics						
Female	0.70	0.72	0.70	0.69	0.77	0.84
	(0.46)	(0.45)	(0.46)	(0.46)		
Year	3.23	3.45	3.26	3.15	0.27	0.49
	(1.42)	(1.18)	(1.48)	(1.42)		
Asian	0.54	0.50	0.54	0.56	0.51	0.79
	(0.50)	(0.51)	(0.50)	(0.50)		
White	0.21	0.20	0.23	0.19	0.83	0.34
	(0.41)	(0.40)	(0.43)	(0.39)		
Ideology	3.01	2.96	3.07	2.97	0.82	0.60
	(1.76)	(1.71)	(1.92)	(1.61)		
Panel B: Private Beliefs						
Renaming Schools	0.39	0.36	0.41	0.38	0.65	0.65
	(0.49)	(0.48)	(0.49)	(0.49)		
Affirmative Action	0.46	0.40	0.45	0.50	0.33	0.35
	(0.50)	(0.49)	(0.50)	(0.50)		
Death Penalty	0.63	0.60	0.65	0.61	0.66	0.50
	(0.48)	(0.49)	(0.48)	(0.49)		
Immunizations	0.82	0.84	0.83	0.81	0.72	0.61
	(0.38)	(0.37)	(0.38)	(0.39)		
DST	0.67	0.74	0.70	0.63	0.28	0.18
	(0.47)	(0.44)	(0.46)	(0.48)		
Panel C: Baseline Guesses						
Renaming Schools	55.60	54.78	55.55	55.89	0.77	0.88
	(21.24)	(22.40)	(20.52)	(21.70)		
Affirmative Action	52.55	50.76	52.73	52.91	0.51	0.94
	(20.80)	(23.95)	(21.14)	(19.52)		
Death Penalty	68.02	68.30	68.03	67.93	0.91	0.96
	(19.48)	(19.46)	(20.09)	(18.98)		
Immunizations	75.19	72.02	76.04	75.29	0.19	0.71
	(18.36)	(19.05)	(18.93)	(17.57)		
DST	63.78	64.86	64.19	63.04	0.74	0.68
	(25.01)	(25.75)	(25.30)	(24.61)		

Notes: This table reports summary statistics from responses from the baseline survey for the whole sample and separately for students randomly assigned to First Movers, Control, and Treatment. P-values from t-tests comparing differences in average values across first and second movers and across Control and Treatment participants are reported in Columns 5 and 6. Year (1-5) indicates number of years in college, and year = 5 for graduate students. Ideology is a 1-7 scale ranging from "extremely liberal" (1) to "extremely conservative" (7). Panel B reports the average private beliefs on each statement, where "0" means disagree "1" means agree. Panel C shows the perceived percentage of Berkeley students who would privately answer "agree" for each topic.

Table C.3: Attrition Table - Total Sample

	Whole Sample (N = 454)		**Completed (N = 383)**		**Attrition (N = 71)**		**T test**
	Mean	Sd	Mean	Sd	Mean	Sd	P Value
Panel A: Demographics							
Female	0.69	0.46	0.70	0.46	0.62	0.49	0.17
Asian	0.53	0.50	0.54	0.50	0.44	0.50	0.10
White	0.22	0.42	0.21	0.41	0.28	0.45	0.19
Year	3.21	1.42	3.26	1.43	2.97	1.35	0.12
Ideology	3.00	1.76	3.01	1.76	2.90	1.76	0.62
Panel B: Private Beliefs							
Rename Schools	0.40	0.49	0.39	0.49	0.46	0.50	0.25
Affirmative Action	0.47	0.50	0.46	0.50	0.52	0.50	0.38
Death Penalty	0.63	0.49	0.63	0.48	0.65	0.51	0.76
Immunizations	0.74	0.44	0.73	0.44	0.79	0.41	0.33
DST	0.68	0.47	0.66	0.47	0.76	0.43	0.10
Panel C: Baseline Guesses							
Rename Schools	55.88	21.31	55.60	21.24	57.38	21.77	0.52
Affirmative Action	52.90	20.64	52.55	20.80	54.75	19.83	0.41
Death Penalty	68.15	19.03	68.02	19.48	68.85	16.53	0.74
Immunizations	70.04	23.56	69.44	24.38	73.27	18.36	0.21
DST	65.01	24.46	64.01	24.98	69.04	21.92	0.12
Panel D: Treatment Assignment							
Treat	0.50	0.50	0.50	0.50	0.49	0.50	0.89

Notes: This table reports summary statistics for the whole sample, the subsample of students who completed all components of the study, and the subsample of students who dropped out of the study. P-values from t-tests comparing differences in average values across those who completed the study versus those who dropped out are reported in Column 7. Year (1-5) indicates number of years in college, and year = 5 for graduate students. Ideology is a 1-7 scale ranging from "extremely liberal" (1) to "extremely conservative" (7). Panel B reports the average private beliefs on each statement where "0" means disagree "1" means agree. Panel C shows the perceived percentage of Berkeley students who would privately answer "agree" for each topic.

Table C.4: Attrition Table - Control Versus Treatment

	Dropouts (N = 71)		**Control Dropouts (N = 33)**		**Treat Dropouts (N = 32)**		**T test**
	Mean	Sd	Mean	Sd	Mean	Sd	P Value
Panel A: Demographics							
Female	0.62	0.49	0.58	0.50	0.62	0.49	0.69
Asian	0.44	0.50	0.42	0.50	0.44	0.50	0.92
White	0.28	0.45	0.33	0.48	0.28	0.46	0.66
Year	2.97	1.35	2.76	1.23	2.94	1.39	0.58
Ideology	2.90	1.76	2.70	1.63	3.12	2.00	0.35
Panel B: Private Beliefs							
Rename Schools	0.46	0.50	0.39	0.50	0.56	0.50	0.17
Affirmative Action	0.52	0.50	0.58	0.50	0.47	0.51	0.40
Death Penalty	0.65	0.51	0.73	0.52	0.56	0.50	0.20
Immunizations	0.79	0.41	0.76	0.44	0.81	0.40	0.60
DST	0.76	0.43	0.82	0.39	0.69	0.47	0.23
Panel C: Baseline Guesses							
Rename Schools	57.38	21.77	52.91	20.72	60.56	21.85	0.15
Affirmative Action	54.75	19.83	52.03	14.11	59.22	22.69	0.13
Death Penalty	68.85	16.53	68.00	14.78	69.66	17.79	0.68
Immunizations	73.27	18.36	73.42	17.23	73.56	19.47	0.98
DST	69.04	21.92	67.91	23.75	69.66	21.02	0.75

Notes: The table reports summary statistics for the students who dropped out of the study in total, as well as separated by their assigned treatment status. P-values from t-tests comparing differences in average values across those who dropped out in the Control group versus those who dropped out in the Treatment group are reported in Column 7. Year (1-5) indicates number of years in college, and year = 5 for graduate students. Ideology is a 1-7 scale ranging from "extremely liberal" (1) to "extremely conservative" (7). Panel B reports the average private beliefs on each statement where "0" means disagree "1" means agree. Panel C shows the perceived percentage of Berkeley students who would privately answer "agree" for each topic.

Table C.5: Summary Statistics of First Movers Used in Information Interventions

	Agree	Disagree	Silent	Total
Renaming Schools	7	5	13	25
Affirmative Action	6	5	13	25
Death Penalty	9	4	12	25
Immunizations	11	1	13	25

Notes: This table reports the number of First Movers who publicly expressed agree, disagree or stayed silent in their Zoom sessions. We presented different versions of these summary statistics to Second Movers depending on their treatment assignment. We used only two sessions to avoid overly-shifting Second Movers' beliefs in the midline survey.

Table C.6: Summary Statistics of All First Movers

	Agree	Disagree	Silent	Total
Renaming Schools	11	9	30	50
Affirmative Action	13	10	27	50
Death Penalty	16	8	26	50
Immunizations	23	2	25	50

Notes: This table reports the number of First Movers who publicly expressed agree, disagree or stayed silent in their Zoom sessions across all four First Mover Zoom sessions. Summary statistics are similar to when we restrict to two randomly selected First Mover sessions in Table C.5.

Table C.7: Public Expression Decisions by Second Movers by Topic

	Logit (Express = 1)			
	(1)	(2)	(3)	(4)
	Affirmative Action	Death Penalty	Immunizations	Renaming Schools
Panel A: Privately Disagree				
Treat	0.111**	0.265***	0.0882	0.210***
	(0.0470)	(0.0864)	(0.119)	(0.0480)
Mean	0.121	0.190	0.214	0.173
SD	0.328	0.395	0.418	0.381
IDs	174	122	60	200
Panel B: Privately Agree				
Treat	-0.499	0.168	0.246	-0.370
	(0.445)	(0.280)	(0.233)	(0.388)
Mean	0.257	0.509	0.423	0.373
SD	0.440	0.502	0.496	0.487
IDs	157	210	272	130
Baseline guesses	✓	✓	✓	✓
Session Controls	✓	✓	✓	✓
Ind Controls	✓	✓	✓	✓

Notes: The outcome variable $Express = 1$ indicates if individual i publicly expresses their true views on topic j. Session controls include time of the session, week of the session, group size and moderator fixed effects. Individual controls include gender, year in school, race and ethnicity, major and self-reported political ideology. Standard errors are clustered at the Zoom session level. Note that standard errors are smaller for the "Privately Disagree" group because the outcome variance is lower in this group, given the lower baseline probability of public expression.* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table C.8: 2SLS: First Stage and Reduced Form

	First Stage				Reduced Form			
	$\%\hat{Agree}^{M}_{i,j}$				Express = 1			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Privately Disagree								
Treat	-7.099***	-6.977***	-7.077***	-7.231***	0.160***	0.161***	0.168***	0.181***
	(1.143)	(1.105)	(1.023)	(1.054)	(0.0320)	(0.0320)	(0.0266)	(0.0309)
Mean	56.49	56.49	56.49	56.49	0.246	0.246	0.246	0.246
SD	16.13	16.13	16.13	16.13	0.431	0.431	0.431	0.431
First-stage F	38.56	39.89	47.87	49.67				
Observations	556	556	556	556	556	556	556	556
Panel B: Privately Agree								
Treat	-6.039***	-6.192***	-6.213***	-5.904***	-0.00965	-0.0114	-0.0105	-0.00122
	(0.995)	(1.149)	(1.090)	(1.000)	(0.0416)	(0.0415)	(0.0403)	(0.0382)
Mean	68.36	68.36	68.36	68.36	0.399	0.399	0.399	0.399
SD	17.37	17.37	17.37	17.37	0.490	0.490	0.490	0.490
First-stage F	36.86	29.06	32.48	37.10				
Observations	769	769	769	769	769	769	769	769
Topic FE	✓	✓	✓	✓	✓	✓	✓	✓
Baseline guesses		✓	✓	✓		✓	✓	✓
Session Controls			✓	✓			✓	✓
Ind Controls				✓				✓

Notes: This table reports the first stage and reduced form regressions corresponding to the 2SLS in Table C.9. The outcome variable $\%\hat{Agree}^{M}_{i_j}$ measures incentivized guesses about the perceived prevalence of the socially appropriate view at midline by individual i for topic j. The outcome variable $Express = 1$ indicates if individual i publicly expresses their true views on topic j. Combined results on four topics (renaming schools affirmative action, death penalty and immunizations) are reported. Session controls include time of the session, week of the session, group size and moderator fixed effects. Individual controls include gender, year in school, race and ethnicity, major and self-reported political ideology. Standard errors are clustered at the Zoom session level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table C.9: 2SLS: Public Expression Decisions by Second Movers

	Express = 1	Express = 1	Express = 1	Express = 1
Panel A: Privately Disagree				
$\%\hat{Agree}^{M}_{i,j}$	-0.0226***	-0.0231***	-0.0237***	-0.0250***
	(0.00549)	(0.00529)	(0.00487)	(0.00542)
Mean	0.164	0.164	0.164	0.164
SD	0.371	0.371	0.371	0.371
IDs	278	278	278	278
Obs	556	556	556	556
Panel B: Privately Agree				
$\%\hat{Agree}^{M}_{i,j}$	0.00160	0.00184	0.00170	0.000207
	(0.00679)	(0.00663)	(0.00634)	(0.00623)
Mean	0.407	0.407	0.407	0.407
SD	0.492	0.492	0.492	0.492
IDs	315	315	315	315
Obs	769	769	769	769
Topic FE	✓	✓	✓	✓
Baseline guesses		✓	✓	✓
Session Controls			✓	✓
Ind Controls				✓

Notes: The outcome variable $Express = 1$ indicates if individual i publicly expresses their true views on topic j. The midline beliefs $\%\hat{Agree}^{M}_{i,j}$ is instrumented by *Treat*. Combined results on four topics (renaming schools affirmative action, death penalty and immunizations) are reported. Session controls include time of the session, week of the session, group size and moderator fixed effects. Individual controls include gender, year in school, race and ethnicity, major and self-reported political ideology. Standard errors are clustered at the Zoom session level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table C.10: Perceived Popularity of the Socially Appropriate View at Midline and Endline

	(1) $\widetilde{\%Agree}$	(2) $\widetilde{\%Agree}$	(3) $\widetilde{\%Agree}$	(4) $\widetilde{\%Agree}$
Treat	0.0251	0.0251	-0.512	-0.496
	(1.159)	(1.159)	(1.090)	(1.114)
Post	5.436***	3.555***	3.555***	3.555***
	(0.889)	(0.719)	(0.721)	(0.722)
Treat X Post	-8.490***	-6.512***	-6.512***	-6.512***
	(1.329)	(1.101)	(1.104)	(1.106)
Endline		3.761***	3.761***	3.761***
		(1.169)	(1.173)	(1.174)
Treat X Endline		-3.955**	-3.955**	-3.955**
		(1.503)	(1.508)	(1.510)
Mean	63.09	63.09	63.09	63.09
SD	22.25	22.25	22.25	22.25
IDs	333	333	333	333
Topic FE	✓	✓	✓	✓
Ind Controls			✓	✓
Session Controls				✓

Notes: The outcome variables are the perceived share of agreement, $\%Agree$, elicited in the baseline, midline, and endline surveys. $Post_t$ is a dummy variable indicating whether the beliefs are elicited after the participant is shown their respective information treatment ($t = 2$ or 3); $Endline_t$ is a dummy variable for whether the beliefs are elicited during the endline survey ($t = 3$); Combined results from the four topics (Affirmative Action, Death Penalty, Immunizations, and Renaming Schools) are reported. Standard errors are clustered at the session level. Individual controls include gender, race and ethnicity, year in school, major, and self-reported political affiliation. Session controls include time and week of the Zoom session, group size, and moderator fixed effects. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table C.11: Private Beliefs Among Silent vs. Expressive Participants

	%Agree Among Silence	%Agree Among Expressed	p-value (one-sided)
Renaming Schools	36.1%	45.6%	0.04
Affirmative Action	46.4%	51.1%	0.22
Death Penalty	54.0%	74.3%	0.00
Immunizations	71.7%	75.5%	0.21
Total	50.9%	64.7%	0.00

Notes: This table reports the share of participants who privately hold the socially appropriate view, comparing those who remained silent with those who publicly expressed their views during the Zoom discussions. P-values from one-sided t-tests of differences between these two samples are reported in Column 3.

Appendix D Supplementary Analysis

D.1 Placebo Topic

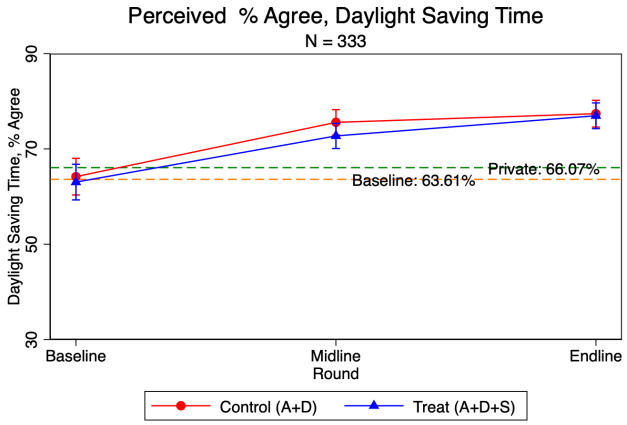


Figure D.3: Perceived %*Agree* Over Time, DST

Note: This graph shows participants' average guesses about the share of agreement, %Agree, for the placebo topic on Daylight Savings Time. Rounds 1, 2 and 3 corresponds to beliefs elicited at the baseline, midline and endline respectively. The dashed green lines show the actual fraction of participants who privately agree with each statement at baseline. The dashed orange line shows the perceived %*Agree* among Control participants at baseline. 95% confidence intervals are included in the graph.

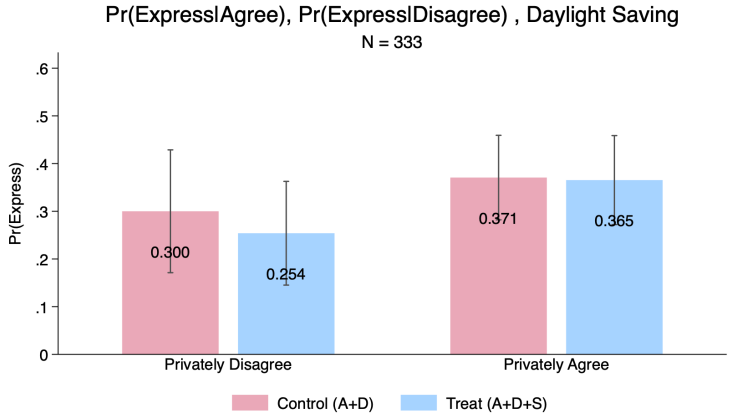


Figure D.4: Public Expression Decisions on Daylight Savings Time (Placebo)

Notes: This graph shows the probability that participants truthfully publicly express their opinions on the placebo topic Daylight Savings Time, separately for participants who privately disagree or privately agree. 95% confidence intervals are included.

Table D.12: Treatment Effects on Inference: DiD Estimates

	Midline Guesses: $\%Agree$			Endline Guesses: $\%Agree$		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	-2.842	-2.426	-2.746	-0.433	-0.00101	1.157
	(2.089)	(1.983)	(2.026)	(3.122)	(2.923)	(3.109)
Controversial Topics	-8.840***	-16.93***	-16.03***	-17.38***	-15.26***	-11.74***
	(1.562)	(1.553)	(1.523)	(2.752)	(2.542)	(2.981)
Treat × Controversial Topics	-3.761*	-4.085**	-4.004**	-10.20***	-10.47***	-9.888***
	(2.139)	(2.019)	(2.021)	(3.090)	(2.861)	(3.321)
Private Agree		4.550***	4.217***		3.857***	2.907***
		(0.834)	(0.848)		(1.033)	(0.964)
$\%\tilde{Agree}^B$		0.120***	0.120***		0.122***	0.0826***
		(0.0221)	(0.0225)		(0.0228)	(0.0208)
Mean	75.57	75.57	75.57	77.39	77.39	77.39
SD	17.53	17.53	17.53	18.47	18.47	18.47
IDs	333	333	333	333	333	333
Topic FE	✓	✓	✓	✓	✓	✓
Baseline guesses		✓	✓		✓	✓
Session Controls						✓
Ind Controls			✓			✓

Notes: The outcome variables are perceived share of agreement, $\%Agree$, elicited in the midline and endline surveys respectively. Controversial Topics = 1 for the Affirmative Action, Death Penalty, Immunizations, and Renaming Schools topics, and = 0 for the Placebo Topic (Daylight Savings Time). Standard errors are clustered at the individual level in columns 1-3 and clustered at the session level in columns 4-6. Session controls include time of the session, week of the session, group size and moderator fixed effects.Individual controls include gender, race and ethnicity, year in school, major, and self-reported political affiliation. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table D.13: Treatment Effects on Public Expression Decisions: DiD Estimates

	OLS (Express = 1)				Logit (Express = 1)
	(1)	(2)	(3)	(4)	(5)
Panel A: Privately Disagree					
Treat	-0.0460	-0.0465	-0.0406	-0.0183	-0.0848
	(0.0877)	(0.0879)	(0.0867)	(0.0902)	(0.470)
Controversial Topics	-0.198***	-0.201***	-0.208***	-0.201***	-1.267***
	(0.0661)	(0.0662)	(0.0668)	(0.0650)	(0.359)
Treat × Controversial Topics	0.206**	0.207**	0.204**	0.186*	1.081**
	(0.0955)	(0.0957)	(0.0956)	(0.0958)	(0.509)
Mean	0.300	0.300	0.300	0.300	0.300
SD	0.463	0.463	0.463	0.463	0.463
IDs	298	298	298	298	298
Obs	669	669	669	669	668
Panel B: Privately Agree					
Treat	-0.00531	-0.00638	-0.00324	-0.0268	-0.125
	(0.0510)	(0.0510)	(0.0486)	(0.0506)	(0.230)
Controversial Topics	-0.143***	-0.127**	-0.128**	-0.151***	-0.774***
	(0.0481)	(0.0548)	(0.0546)	(0.0506)	(0.261)
Treat × Controversial Topics	-0.00435	-0.00440	-0.00526	0.0141	0.0738
	(0.0534)	(0.0538)	(0.0514)	(0.0511)	(0.232)
Mean	0.371	0.371	0.371	0.371	0.371
SD	0.485	0.485	0.485	0.485	0.485
IDs	330	330	330	330	330
Obs	989	989	989	989	989
Topic FE	✓	✓	✓	✓	✓
Baseline guesses		✓	✓	✓	✓
Session Controls			✓	✓	✓
Ind Controls				✓	✓

Notes: The outcome variable $Express = 1$ indicates if individual i publicly expresses their true views on topic j. Controversial Topics = 1 for the Affirmative Action, Death Penalty, Immunizations, and Renaming Schools topics, and = 0 for the Placebo Topic (Daylight Savings Time). Session controls include time of the session, week of the session, group size and moderator fixed effects. Individual controls include gender, race and ethnicity, year in school, major and self-reported political ideology. Standard errors are clustered at the Zoom session level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

D.2 Attention and Recall

Table D.14: Heterogeneous Treatment Effects on Perceived %*Agree*, Midline and Endline

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Midline									
Treat	-7.336***	-6.394***	-7.065***	-8.591***	-6.209***	-6.626***	-6.446***	-7.038***	-10.64***
	(1.357)	(1.383)	(1.752)	(1.418)	(1.070)	(1.862)	(1.282)	(1.161)	(3.293)
Treat x Private Agree	1.084								1.066
	(1.528)								(1.597)
Treat x Soph1		-0.472							-0.364
		(1.646)							(1.668)
Treat x Soph2		-2.325							
		(1.567)							
Treat x Female			0.506						0.771
			(2.040)						(2.043)
Treat x Asian				3.335*					4.097*
				(1.912)					(2.309)
Treat x White					-2.324				0.405
					(2.338)				(2.804)
Treat x Liberal						-0.105			0.321
						(2.146)			(2.225)
Treat x Senior							-0.559		-0.363
							(1.909)		(1.874)
Treat x Social Science								1.287	2.120
								(1.940)	(2.011)
Mean	66.73	66.73	66.73	66.73	66.73	66.73	66.73	66.73	66.73
SD	17.11	17.11	17.11	17.11	17.11	17.11	17.11	17.11	17.11
IDs	333	333	333	333	333	333	333	333	333
Panel B: Endline									
Treat	-13.96***	-7.951***	-7.980***	-11.82***	-9.938***	-10.70***	-10.88***	-10.00***	-14.29***
	(1.284)	(1.414)	(1.243)	(1.550)	(1.259)	(2.230)	(1.381)	(1.246)	(3.527)
Treat x Private Agree	6.099***								4.884***
	(1.706)								(1.581)
Treat x Soph1		-3.198*							-2.634
		(1.642)							(1.671)
Treat x Soph2		-3.875**							
		(1.710)							
Treat x Female			1.225						1.227
			(2.097)						(1.940)
Treat x Asian				2.499					3.071
				(1.758)					(2.480)
Treat x White					-2.201				-0.117
					(2.067)				(2.775)
Treat x Liberal						0.409			-0.0544
						(2.787)			(2.350)
Treat x Senior							1.063		2.100
							(1.921)		(1.700)
Treat x Social Science								-1.530	-1.499
								(1.605)	(1.480)
Mean	70.45	70.45	70.45	70.45	70.45	70.45	70.45	70.45	70.45
SD	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00	18.00
IDs	333	333	333	333	333	333	333	333	333

Notes: The outcome variables are the perceived share of agreement, %*Agree*, elicited in the midline survey (Panel A) and the endline survey (Panel B). Combined results on the four topics (renaming schools, affirmative action, death penalty, and immunizations) are reported. We classify a respondent as sophisticated ("Soph1") if they correctly guess that the share of disagreement is higher among the silent subsample than among the expressed subsample. We also report results for an alternative definition of sophisticated ("Soph2") if they correctly guess the empirical share of silent participants who privately disagree for each topic, rounded to the nearest 10th percentile. Standard errors are clustered at the individual level in Panel A and clustered at the session level in Panel B. Session controls include time of the session, week of the session, group size and moderator fixed effects. Individual controls include gender, year in school, race and ethnicity, major and self-reported political ideology. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

D.3 Multi-Modal Beliefs

We conduct supplementary analysis to explore the possibility that beliefs are multi-modal, which helps us to identify the mechanisms underlying our experimental results when people may begin with different priors. In our study context, this would imply that some people attend to silence and some do not, leading to bi-modal beliefs. Then our Treatment intervention could be interpreted as potentially shifting the share of individuals who attend to silence.

If such bi-modality is prevalent, it should be most evident in midline beliefs, when participants are exposed to the same information about the first movers. We thus examine belief distributions using three complementary approaches. First, we plot the kernel density of the belief distribution in the midline to see if there is a multi-modal pattern. Second, we estimate two-type finite mixture models to formally test for the presence of latent types that differ in their attention to silence and to explore whether the treatment shifts the composition of these types. Finally, we link the latent types identified in the mixture model to participants' recall of silence in the Zoom sessions to check whether the silence-attentive type corresponds to those who correctly remember silence in Zoom sessions.

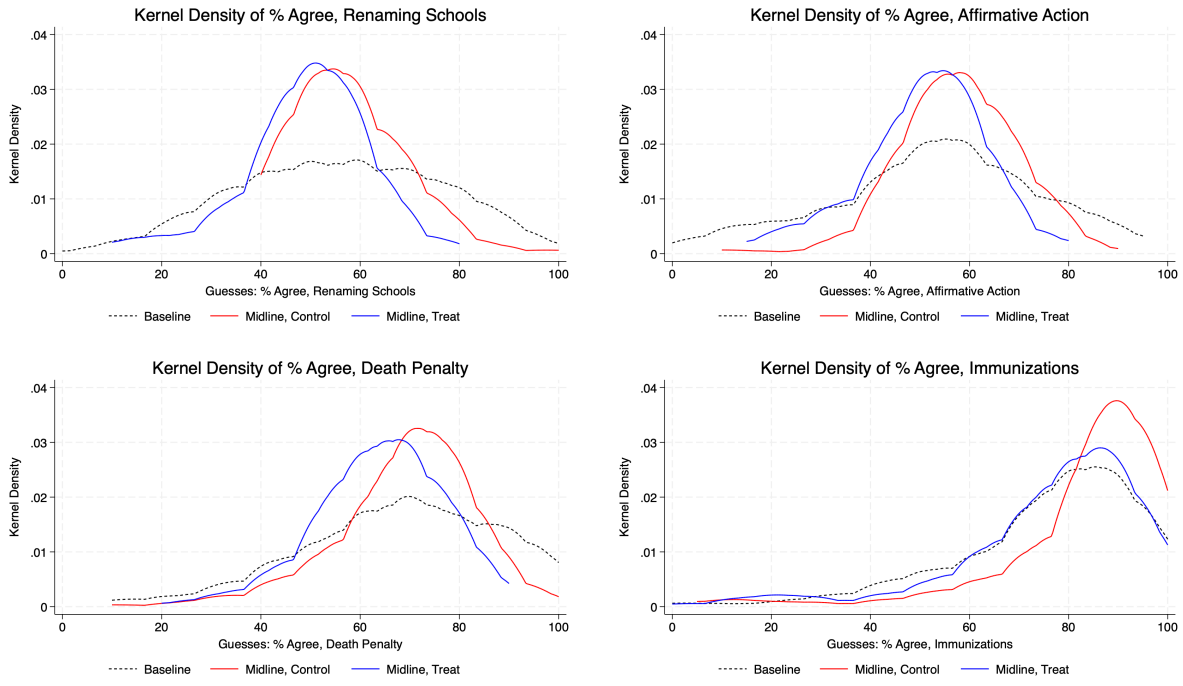


Figure D.5: Kernel Density of Perceived $\%Agree$

Note: These graphs show the kernel density of participants' guesses about the share of agreement for each statement, %Agree. We report beliefs elicited at baseline and midline (after the information treatment and before the Zoom session), for treatment and control groups respectively.

Figure D.5 plots kernel densities of the perceived share of agreement across the four

main topics. While the distributions are not cleanly bimodal, the information treatment systematically re-weights probability mass towards the lower-belief region across three of the four topics (a left shoulder in treated density). This pattern is consistent with a mixture-of-types interpretation, in which some participants shift into a silence-attentive type under treatment.

Next, we formally examine the belief distribution in the midline using a finite mixture model. A two-component normal mixture provides a better fit to the empirical distribution compared to a single normal (ΔBIC is -19.2, -9.3 and -11.1 respectively), consistent with the idea that beliefs are a mixture of two types.

Table D.15: Treatment Effects on Latent Attention Types Across Topics

Topic	Renaming Schools	Affirmative Action	Death Penalty	Immunizations
%*Agree*, Type 1	38.87	41.13	60.50	63.16
	(5.91)	(5.49)	(2.73)	(2.83)
%*Agree*, Type 2	55.54	57.44	73.72	88.62
	(0.82)	(0.81)	(1.43)	(0.60)
Pr(Type 1| Control)	0.00	0.05	0.35	0.23
	(0.00)	(0.04)	(0.12)	(0.04)
Pr(Type 1| Treat)	0.28	0.34	0.72	0.48
	(0.15)	(0.15)	(0.17)	(0.06)
p-value	0.058	0.021	0.00	0.00

Notes: This table reports topic-specific estimates from two-component finite mixture models of midline beliefs about the perceived share of agreement. For each topic, "Type 1" and "Type 2" correspond to the two latent belief types (we interpret as silence-attentive and inattentive types respectively), with rows "%*Agree*, Type 1" and "%*Agree*, Type 2" reporting the estimated mean perceived agreement within each type (in percentage points). "Pr(Type 1 | Control)" and "Pr(Type 1 | Treat)" report the predicted probabilities of belonging to Type 1 for the control and treatment participants, respectively. P-values from a test of whether treatment significantly affects Type-1 membership in the latent-class probability equation are included. Standard errors for predicted probabilities are computed using the delta method.

Table D.15 reports results from topic-specific two-type mixture models of perceived agreement. Across all four topics, the two latent types differ in their average beliefs about the share of agreement, with Type 1 having significantly lower estimates than Type 2. We tend to interpret Type 1 as those who pay attention to silence. Furthermore, the treatment, which is intended to increase attention to silence, shifts the composition of these latent types. For Renaming Schools and Affirmative Action, the probability of belonging to Type 1 (the attentive type) increases from near zero in the control group to 0.28 and 0.34 under treatment, respectively. For Death Penalty and Immunizations, the corresponding increases are 0.37 and 0.25 respectively. These results indicate that the information intervention does not merely shift beliefs uniformly, but instead increases the share of participants who adopt a lower-belief, silence-attentive type.

To validate the interpretation of the silence-attentive type, we also link participants' type to their recall of silence in Zoom sessions. For each participant, we first generate individual-

level type predictions from the estimated two-component mixture model by computing the posterior probability of belonging to each latent type given their observed guesses. We classify a participant as the silence-attentive type (Type 1) if their posterior probability of belonging to the low-belief type exceeds 0.5. As discussed in Section 5.2, we interpret a respondent's accuracy of recall about the Zoom sessions they attended as a proxy measure for the level of attention they paid during the Zoom discussions. We find a positive and statistically significant correlation between correct recall and classification as the low-belief, silence-attentive type (Type 1) ($\rho = 0.072$, p = 0.009). This indicates that participants who remember the silence information are systematically more likely to be identified as silence-attentive in the mixture model.

D.4 Intrinsic Value of Expression

In our conceptual framework, we distinguish between the intrinsic value of expression versus perceived social backlash as key factors in the decision to express one's views. Empirically, we can also separately identify the effects of these two forces in observed expression decisions. To proxy for the intrinsic value of expression, we use participants' public expression decisions on the placebo topic, Daylight Savings Time (DST). DST is a socially neutral topic, as validated by our Krupka and Weber exercise, and is plausibly orthogonal to other controversial topics discussed in the Zoom sessions. We therefore interpret a participant's willingness to express on this placebo topic as capturing an individual's intrinsic propensity to speak up.

Table D.16 shows that this proxy is strongly correlated with expression on the four other topics: individuals classified as having a high intrinsic value of expression are significantly more likely to express their views, both when they privately agree and disagree with the socially appropriate view. Importantly, even after controlling for this proxy, we still find a large and significant treatment effects on expression among participants who privately disagree. Combined with our 2SLS results (Tables C.8 and C.9), this pattern suggests that the treatment operates through shifting beliefs about the prevalence of socially appropriate views, thereby reducing the perceived social sanction costs for expressing dissent. Taken together, these results show that increasing the salience of silence significantly changes both inference and expression, separate from individual differences in the intrinsic propensity of speaking up.

Table D.16: Treatment Effects on Expression, Controlling for Intrinsic Value of Expression

	Express = 1, Privately Disagree		Express = 1, Privately Agree	
	(1)	(2)	(3)	(4)
Treat	0.161***	0.170***	-0.00196	-0.0104
	(0.0301)	(0.0282)	(0.0369)	(0.0339)
High V (Expressed on DST)	0.111***	0.0753*	0.203***	0.191***
	(0.0407)	(0.0373)	(0.0390)	(0.0394)
Mean	0.164	0.164	0.407	0.407
SD	0.371	0.371	0.492	0.492
IDs	298	298	330	330
Obs	556	556	769	769
Topic FE	✓	✓	✓	✓
Baseline guesses	✓	✓	✓	✓
Session Controls		✓		✓
Ind Controls		✓		✓

Notes: The outcome variable $Express = 1$ indicates if individual i publicly express their true views on topic j. We use individual i's expression decision on the placebo topic (Daylight Savings Time) as a proxy for their individual-level intrinsic value of expression. Combined results on four topics (renaming schools, affirmative action, death penalty, and immunizations) are reported. Session controls include time of the session, week of the session, group size and moderator fixed effects. Individual controls include gender, race and ethnicity, year in school, major and self-reported political ideology. Standard errors are clustered at the Zoom session level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

D.5 Spillover Effects

In the endline survey, we asked students if they would be willing to publicly share their views on a sensitive topic that was not discussed in the Zoom sessions to measure potential treatment spillover effects. Specifically, participants were reminded of a statement they had previously evaluated: *"Transgender athletes should not be allowed to compete in female sports at the college level."* They were then asked whether they would like to publicly share their opinion on the topic with other Berkeley students participating in the study, and were given a text box to briefly explain their reasoning. Participants also had the option to skip the question (stay silent).

Figure D.6 shows evidence for moderate spillover effects: among participants who privately hold the socially inappropriate view ("privately disagree"), those assigned to the treatment group are about 15 percentage points (60%) more likely to publicly express their views on this additional topic than those in the control group (p-value = 0.04). This suggests that our information treatment may have broader effects beyond the topics covered in the treatments themselves or in the Zoom sessions.

That said, we interpret these results with caution. The spillover effects likely depend on how participants perceive the correlation between topics. In our setting, some students may view the transgender-athletes topic as related to broader debates about identity or fairness, which makes it plausible that the treatment effects on other related topics might affect

their expression decisions about this particular topic. Indeed, it appears that views on the transgender athlete topic are positively correlated with views on other topics included in the information treatment and Zoom sessions, notably the Affirmative Action and Death Penalty topics (see Figure D.7). However, such perceived correlations are inherently topic-specific. For issues that participants do not see as correlated, we would not necessarily expect similar cross-topic effects. Therefore, we interpret the observed spillover effects as suggestive.

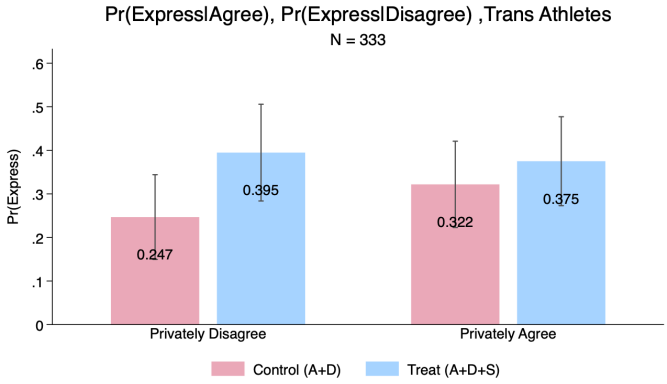


Figure D.6: Public Expression Decisions on Transgender Athletes Topic

Note: This graph shows participants' public expression decisions in the endline survey, separately by private disagreement or agreement, for a topic not included in the information treatments or Zoom sessions (Transgender athletes) to measure potential treatment spillover effects. 95% confidence intervals are included.

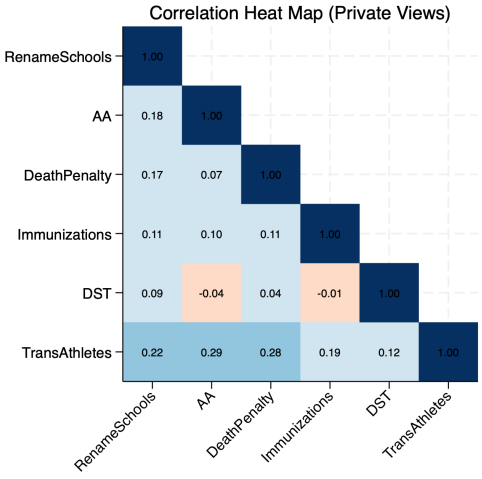


Figure D.7: Correlation: Private Views on Different Topics

Note: This figure displays the correlation coefficients across the five topics included in the information treatments and Zoom discussions - Renaming schools, affirmative action, death penalty, immunizations, daylight savings time (placebo) - as well as an additional topic only included in the baseline and endline surveys, transgender athletes.

Appendix E Extensions of Main Experiment

E.1 Silence and Polarization

We use a two-stage survey to elicit private beliefs, public expressions and perceived belief distribution, on the same four topics that we tested in the main experiment.

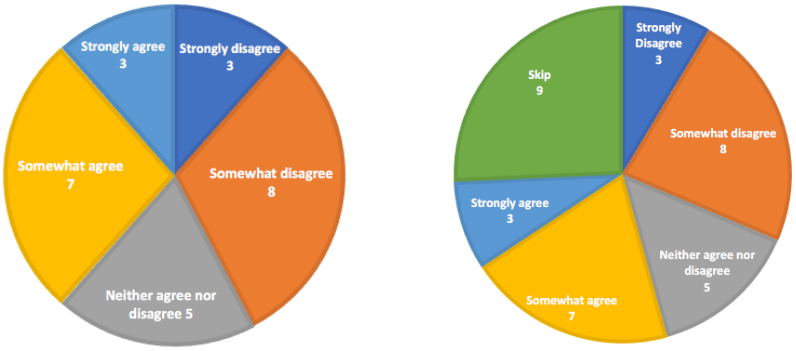
We first elicit private beliefs and guesses about other students' views in a baseline survey. Instead of a binary "agree" or "disagree" option, participants are provided with a 5-point Likert scale, ranging from "strongly agree" to "strongly disagree". They are then incentivized to guess, "Among other Berkeley students who participate in this study, what fraction do you think would privately answer 'strongly agree', 'somewhat agree', 'neither agree nor disagree', 'somewhat disagree', 'strongly disagree' respectively?" for each statement.

Following the same procedure as the original experiment, we randomly assign participants who complete the baseline survey into a First Movers group, Control group, or Treatment group. In the follow-up survey, which is sent one week after the baseline survey, we elicit First Movers' public expression decisions. Participants are asked to share their opinions on a subset of topics, where their survey responses will be shared publicly with all other study participants. Participants also have the option to stay silent, or not share their opinion, for each topic. Specifically, they are informed that: *"Your answers will be shared publicly with other students participating in the study. Feel free to skip the questions if you don't want to answer it or if you don't want to share your answers with other students."* A key difference from the main experiment is that students can choose their response from the 5-point scale rather than the original binary scale.

As in the main experiment, we randomly vary the salience of silence when presenting summary statistics about First Movers' public expression to Second Movers in the follow-up survey. The number of students who skipped the public expression question for each topic is highlighted in the Treatment group information but not in the Control group information (Figure E.8). Second Movers then guess, incentivized for accuracy, the percentage of Berkeley participants who would privately choose each option, ranging from "strongly agree" to "strongly disagree." They are also asked to provide their own opinions on the same 5-point scale which are made public to the other study participants.

Expression by First Movers

The results from this extension show that public expression is selected not only based on the socially appropriateness of opinions but also the strength of opinions. Table E.17 presents the expression decisions by First Movers. The outcome variable is a dummy variable that is equal to one if participant i publicly expresses their opinion on topic j in the survey. Participants



Polarization Control Condition Polarization Treatment Condition

Figure E.8: Polarization Extension: Information Treatments

Note: This figure shows an example of the information interventions shared with Control and Treatment group participants respectively for a replication that allows opinions to be expressed on a 5-point scale, ranging from "Strongly Disagree" to "Strongly Agree."

holding socially acceptable views "agree" (answering "Somewhat Agree" or "Strongly Agree" in the baseline survey) are 44% (27.1 percentage points) more likely to publicly express their views on average. Moreover, participants holding extreme views (answering either "Strongly Agree" or "Strongly Disagree") are around 33% (20 percentage points) more likely to publicly express them on average. Taken together, public expression is skewed in the direction of socially appropriate views and is more heavy-tailed than the actual belief distribution.

Inference by Second Movers

The results on inference replicate our main results and furthermore show that experimentally drawing attention to silence also reduces perceived polarization. Table E.18 shows that Treatment participants have lower estimates of the popularity of socially appropriate views relative to Control participants. Moreover, treatment participants also have lower guesses about the prevalence of extreme views, suggesting that increased attention to silence leads to lower perceived polarization.

Expression by Second Movers

Finally, we also elicit Second Movers' willingness to publicly express their opinions and find similar dynamic patterns as in our main experiment. Table E.19 demonstrates the public expression decisions by Second Movers who privately disagree with the socially appropriate views (answering "Strongly Disagree" or "Somewhat Disagree" in the baseline survey), and who privately hold moderate views (answering "Somewhat Disagree," "Neither Agree nor Disagree," or "Somewhat Agree"). The positive estimated coefficient for *Treat* suggests that those with socially inappropriate views or moderate views are more likely to publicly express their views in the treatment group relative to the control group.

Table E.17: Polarization Extension: Expression Decisions by First Movers

	(1) Express =1	(2) Express =1	(3) Express = 1
Private Agree	0.334***	0.300***	0.266**
	(0.0905)	(0.0908)	(0.100)
Private Extreme	0.229***	0.220***	0.205**
	(0.0822)	(0.0790)	(0.0872)
Mean	0.629	0.629	0.629
SD	0.485	0.485	0.485
IDs	35	35	35

Notes: This table reports the public expression decisions by First Movers on four topics (Affirmative Action, Death Penalty, Immunizations, and Renaming Schools). $Express = 1$ indicates if individual i publicly expresses their opinions on topic j. Private Agree indicates a respondent privately responded either "Somewhat Agree" or "Strongly Agree." Private Extreme indicates a respondent privately responded either "Strongly Agree" or "Strongly Disagree." Demographic controls include gender, race and ethnicity, year in school, major, and self-reported political ideology. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table E.18: Polarization Extension: Treatment Effects on Inference

	(1) $\%Agree$	(2) $\%Agree$	(3) $\%Disagree$	(4) $\%Disagree$	(5) $\%Extreme$	(6) $\%Extreme$
Treat	-4.175***	-3.535**	3.036***	2.654**	-3.240**	-1.834
	(1.334)	(1.411)	(1.132)	(1.175)	(1.443)	(1.359)
Topic FE	✓	✓	✓	✓	✓	✓
Baseline guesses		✓		✓		✓
Demographics		✓		✓		✓
Mean	52.74	52.74	29.20	29.20	35.85	35.85
SD	19.56	19.56	14.13	14.13	15.48	15.48
IDs	123	123	123	123	123	123

Notes: This table reports Second Movers' inference about the belief distribution in an extension of the main experiment that measures views using a 5-point Likert scale. The outcome variables include the perceived share of agreement, %Agree, pooling together "Somewhat Agree" and "Strongly Agree" responses; the perceived share of disagreement, %Disagree, pooling together "Somewhat Disagree" and "Strongly Disagree" responses; and the perceived share of extreme views, %Extreme, pooling together "Strongly Agree" and "Strongly Disagree" responses. Pooled results across four topics (Affirmative Action, Death Penalty, Immunizations, and Renaming Schools) are reported. Demographic controls include gender, race and ethnicity, year in school, major, and self-reported political ideology. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table E.19: Polarization Extension: Treatment Effects on Expression

	Private Disagree		Private Moderate	
	Express = 1	Express = 1	Express = 1	Express = 1
Treat	0.124	0.189*	0.131	0.228**
	(0.0853)	(0.102)	(0.0840)	(0.0878)
Topic FE	✓	✓	✓	✓
Baseline guesses		✓		✓
Demographics		✓		✓
Mean	0.494	0.494	0.518	0.518
SD	0.501	0.501	0.500	0.500
IDs	123	123	123	123

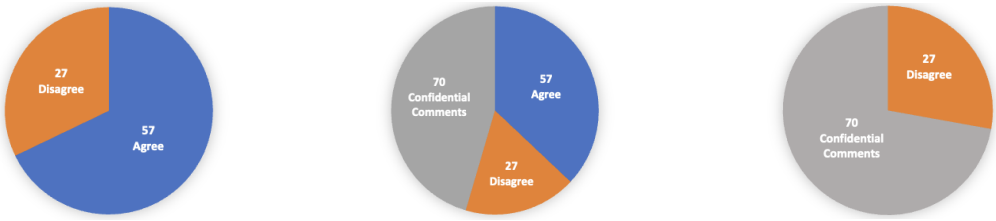
Notes: This table reports the public expression decisions by Second Movers on four topics (Affirmative Action, Death Penalty, Immunizations, and Renaming Schools). $Express = 1$ if individual i expresses their opinion on topic j. Private Disagree indicates a respondent privately responded either "Somewhat Disagree" or "Strongly Disagree." Private Moderate indicates a respondent privately responded either "Somewhat Disagree," "Neither Agree nor Disagree," or "Somewhat Agree." Demographic controls include gender, race and ethnicity, year in school, major, and self-reported political ideology.* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

E.2 Berkeley Building Name Review Committee

The renaming of school buildings is one of the topics that was discussed in our Zoom experiment, and it naturally extends to another experiment with Berkeley students. In this follow-up experiment, we conducted a brief survey with Berkeley students to understand their private views, perceptions about others' views, and public expression decisions on renaming Berkeley buildings. We first elicit students' private views. Participants are asked if they agree or disagree with the following statement: *"All Berkeley buildings named after controversial historical figures should be renamed."*

We then share summaries of this public discussion with students, where we randomly vary which version of the information intervention students receive. In total, 255 participants were randomly assigned to one of three groups of equal size: 1) Control, which omits the number of confidential comments; 2) Treatment, which includes the number of all three types of comments; or 3) Omit Agree, which omits the number of public agreement comments in the summary pie charts that are shared with participants. An example of these information interventions are shown in Figure E.9.

After presenting the information interventions, we incentivize students to guess the belief distribution by asking them *"Among other Berkeley students who participate in this survey, what percentage do you think privately 'agree" and 'disagree" with the following statement: 'All Berkeley buildings named after controversial historical figures should be renamed'".*



Control Treatment Omit Agree

Figure E.9: Experimental Treatments Varying Omitted Information

Notes: This figure shows the information interventions shared in an extension of the main experiment with Control, Treatment, and Omit Agree participants respectively, summarizing the community comments provided for a proposed renaming of Moses Hall on UC Berkeley's campus.

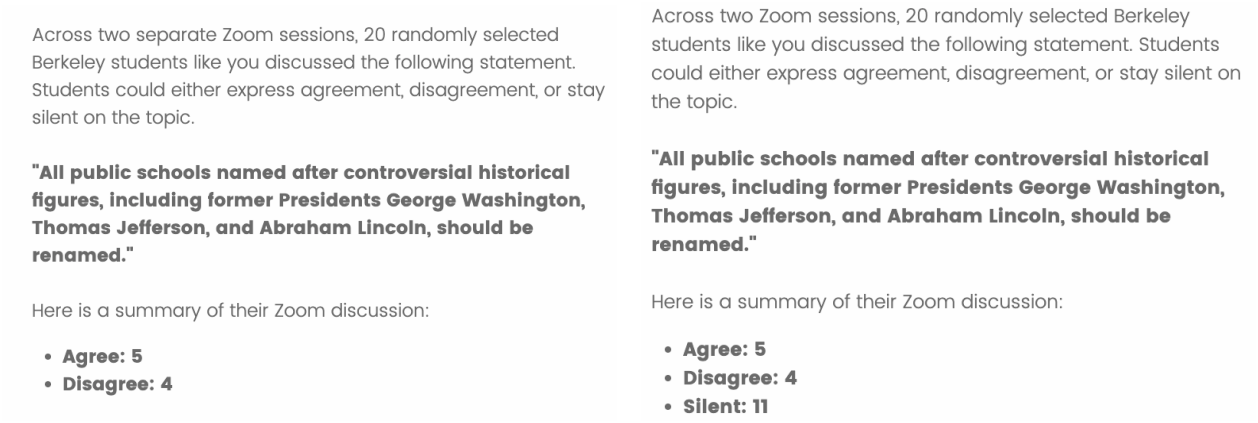
Finally, we elicit participants' public expression decisions by asking them to sign a petition that is either in favor of or against the renaming of other Berkeley buildings. The petition that participants receive depends on their private beliefs. For example, those who privately agree with renaming are presented with a petition titled "Don't Stop at Moses: Remove ALL Racist Names from UC Berkeley NOW", and those who privately disagree receive a petition titled 'Please Don't Erase our History in UC Berkeley." Participants are informed that the petitions are likely to be shared with the Berkeley Building Name Review Committee and, if they choose to sign the petition, they must do so with their real names and Berkeley email addresses. Results on inference and expression are reported in Table **??** below.

Table E.20: Renaming Committee Extension: Treatment Effects on Inference and Expression

	Guesses, %Agree		Pr(Sign|Disagree)		Pr(Sign|Agree)	
	(1)	(2)	(3)	(4)	(5)	(6)
Omit Agree	-7.707***	-6.933***	0.132*	0.117	-0.0402	0.0333
	(2.601)	(2.620)	(0.0761)	(0.0857)	(0.109)	(0.127)
Treatment	-8.963***	-8.094***	0.205***	0.226**	-0.0152	0.0259
	(2.609)	(2.598)	(0.0765)	(0.0862)	(0.109)	(0.126)
p-value Omit Agree - Treatment	0.631	0.655	0.344	0.206	0.820	0.953
Mean	67.58	67.58	0.0455	0.0455	0.390	0.390
SD	15.30	15.30	0.211	0.211	0.494	0.494
IDs	254	254	254	254	254	254
Individual Controls		✓		✓		✓

Notes: The outcome variable in columns 1-2 is participants' guesses about the share of agreement among Berkeley students. Columns 3-6 report the willingness to sign a petition either against or in favor of renaming, where the petition that participants receive corresponds to their private beliefs. The omitted group is the Control group. Individual Controls include gender, race and ethnicity, year in school, major, and self-reported political ideology. The p-value from testing equality of the estimated treatment effects of the Omit Agree and Treatment interventions are included. Standard errors are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

E.3 Replication Experiment



New Control Condition New Treatment Condition

Figure E.10: Revised Information Treatments for Replication

Note: This figure shows an example of the revised information interventions shared with Control and Treatment group participants respectively, summarizing the discussions of the First Movers of the Renaming Schools topic, for a replication testing the robustness of the original experimental results to variations in the design of the information interventions.

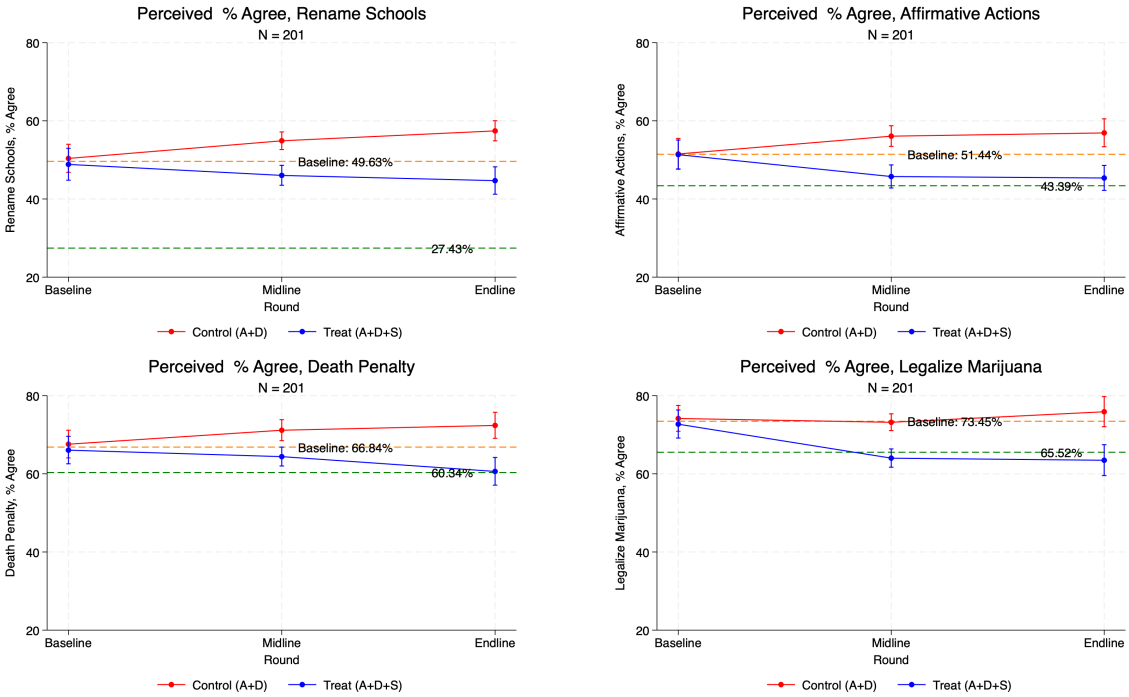


Figure E.11: Replication: Perceived %*Agree* Over Time

Note: These graphs show participants' average guesses about the share of agreement for each statement, %Agree, in a replication experiment testing the robustness of the original experimental results to variations in the design of the information treatments. We report beliefs elicited at baseline, midline (after the information treatment and before the Zoom session), and endline (after the Zoom session) respectively. The dashed green lines show the fraction of participants who privately agree with each statement at baseline. The dashed orange line shows the perceived %*Agree* among Control participants in the baseline. 95% confidence intervals are included.

Table E.21: Replication: Treatment Effects on Inference

	Midline Guesses: $\%Agree$			Endline Guesses: $\%Agree$		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	-8.645***	-8.346***	-8.574***	-12.02***	-11.47***	-13.60***
	(1.251)	(1.196)	(1.205)	(3.018)	(2.932)	(2.125)
Private Agree		4.729***	3.761***		5.567***	4.531***
		(1.208)	(1.312)		(1.378)	(1.536)
$\%\tilde{Agree}^B$		0.124***	0.123***		0.0471*	0.0246
		(0.0292)	(0.0299)		(0.0269)	(0.0302)
Mean	60.71	60.71	60.71	62.25	62.25	62.25
SD	14.88	14.88	14.88	17.76	17.76	17.76
IDs	201	201	201	201	201	201
Obs	603	594	594	603	594	594
Topic FE	✓	✓	✓	✓	✓	✓
Baseline guesses		✓	✓		✓	✓
Session Controls						✓
Ind Controls			✓			✓

Notes: This table reports treatment effects on inference in a replication experiment testing the robustness of the main experimental results to variations in the design of the information treatments. The outcome variables are the perceived share of agreement, $\%Agree$, elicited in the midline and endline surveys respectively. Combined results from the four topics (Affirmative Action, Death Penalty, Renaming Schools, and Legalizing Marijuana) are reported. Standard errors are clustered at the individual level in columns 1-3, and clustered at the session level in columns 4-6. Session controls include time of the session, week of the session, group size and moderator fixed effects. Individual controls include gender, race and ethnicity, year in school, major, and self-reported political affiliation. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table E.22: Replication: Treatment Effects on Expression

	OLS (Express = 1)				Logit (Express = 1)
	(1)	(2)	(3)	(4)	(5)
Panel A: Privately Disagree					
Treat	0.139**	0.141**	0.197***	0.172***	0.175***
	(0.0565)	(0.0579)	(0.0318)	(0.0362)	(0.0342)
Mean	0.191	0.191	0.191	0.191	0.191
SD	0.395	0.395	0.395	0.395	0.395
IDs	179	179	179	179	179
Obs	401	401	401	401	399
Panel B: Privately Agree					
Treat	-0.0261	-0.0266	-0.0148	0.0123	0.00846
	(0.0447)	(0.0448)	(0.0447)	(0.0364)	(0.0348)
Mean	0.442	0.442	0.442	0.442	0.442
SD	0.498	0.498	0.498	0.498	0.498
IDs	177	177	177	177	177
Obs	392	392	392	392	392
Topic FE	✓	✓	✓	✓	✓
Baseline guesses		✓	✓	✓	✓
Session Controls			✓	✓	✓
Ind Controls				✓	✓

Notes: This table reports treatment effects on public expression decisions in a replication experiment testing the robustness of the main experimental results to variations in the design of the information treatments.The outcome variable $Express = 1$ indicates whether individual i publicly expresses their true views on topic j. Combined results on four topics (renaming schools, affirmative action, death penalty and legalizing marijuana) are reported. Session controls include time of the session, week of the session, group size and moderator fixed effects. Individual controls include gender, race and ethnicity, year in school, major and self-reported political ideology. Standard errors are clustered at the Zoom session level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

E.4 ANES Data

Survey Questions

1. In politics people sometimes talk of left and right.

2. We hear a lot of talk these days about liberals and conservatives. Here a seven-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative.

3. Some people think the government should provide fewer services, even in areas such as health and education, in order to reduce spending. Suppose these people are at one end of a scale, at point 1. Other people feel that it is important for the government to provide many more services even if it means an increase in spending. Suppose these people are at the other end, at point 7.

4. Some people believe that we should spend much less money for defense. Suppose these people are at one end of a scale, at point 1. Others feel that defense spending should be greatly increased. Suppose these people are at the other end, at point 7.

5. There is much concern about the rapid rise in medical and hospital costs. Some feel there should be a government insurance plan which would cover all medical and hospital expenses for everyone. Suppose these people are at one end of a scale, at point 1. Others feel that all medical expenses should be paid by individuals, and through private insurance plans like Blue Cross or other company paid plans. Suppose these people are at the other end, at point 7.

6. Some people feel that the government in Washington should see to it that every person has a job and a good standard of living. Suppose these people are at one end of a scale, at point 1. Others think the government should just let each person get ahead on his/their own. Suppose these people are at the other end, at point 7.

7. Some people feel that the government in Washington should make every effort to improve the social and economic position of blacks. Suppose these people are at one end of a scale, at point 1. Others feel that the government should not make any special effort to help blacks because they should help themselves. Suppose these people are at the other end, at point 7.

Questions: for all of the topics above, ANES asks:

- Where would you place yourself on this scale?

- Where would you place the Democratic Party?
- Where would you place the Republican Party?

Table E.23: ANES: Silence and Misperceptions

	Misperception (Raw)	Misperception (Raw)	Misperception (Raw)	Misperception (Adjusted)	Misperception (Adjusted)	Misperception (Adjusted)
Panel A: All Sample						
% No Response	0.0506	0.278***	0.169**	0.141**	0.406***	0.194*
	(0.0481)	(0.0737)	(0.0785)	(0.0633)	(0.108)	(0.112)
Mean	0.0568	0.0568	0.0568	0.0410	0.0410	0.0410
SD	0.0448	0.0448	0.0448	0.0596	0.0596	0.0596
N	184	184	184	184	184	184
Topic Salience		✓	✓		✓	✓
Year & Topic FE		✓	✓		✓	✓
Actual Beliefs			✓			✓
Panel B: In-Person Surveys						
% No Response	0.0220	0.223***	0.154*	0.105*	0.397***	0.204*
	(0.0470)	(0.0750)	(0.0816)	(0.0621)	(0.108)	(0.118)
Mean	0.0583	0.0583	0.0583	0.0422	0.0422	0.0422
SD	0.0459	0.0459	0.0459	0.0611	0.0611	0.0611
N	184	184	184	184	184	184
Topic Salience		✓	✓		✓	✓
Year & Topic FE		✓	✓		✓	✓
Actual Beliefs			✓			✓

Notes: *Non-response* is a dummy variable that is equal to 1 if participants choose either of the following survey responses for each topic: *Refused; Don't know; or Haven't thought much about this.* Misperception (Raw) takes the absolute differences between the actual ratings by Democrats or Republicans and the perceived ratings by all respondents. Misperception (Adjusted) standardizes the direction so that positive misperceptions are consistent with party stereotypes. Results for the entire sample of ANES data are reported in Panel A and for the subset of ANES data when surveys were conducted in-person are reported in Panel B. Standard errors are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix F Survey Instructions

F.1 Recruitment + Baseline Survey

[**Welcome**] You are invited to participate in a research study. The purpose of the study is to understand college students' views on some political and socioeconomic issues.

To participate in the study, you will attend one Zoom session and complete three short surveys. The study will take a total of 60 minutes. As appreciation for your time, you will be compensated a total of $20-$24 for your participation, which will be distributed to you as an Amazon gift card after completing ALL parts of the study. During the study:

1. You will answer this survey, which should take about 7 minutes to complete. This survey will include questions about various political and socioeconomic issues. Your survey responses will be completely private and observable only to researchers.

2. You will take part in a Zoom session with other study participants that will last 45 minutes. During the Zoom session, you will discuss a set of political and socioeconomic issues with other study participants. The discussion sessions will take place between 3-5pm PT Monday - Friday. You need to attend ONE of the sessions, and you can choose the session that best fits your schedule. The slots will be posted on SONA, please check your email inbox for a notification when the slots are posted.

3. Immediately after the Zoom session, you will answer a brief survey, which should take about 5 minutes to complete. Again your survey responses will be private and observable only to researchers.

Participation in this study is completely voluntary. You have the right to decline to participate or to withdraw at any point in this study. To learn more details about the study, please review the following consent document: [link to the consent file]

[**Private Beliefs**] In the first part of this survey, you will read statements on 10 political and socioeconomic issues. After reading each statement, you will be asked whether you agree or disagree with the statement. Please note that you can skip any question that you do not want to answer. Your survey responses will be kept private.

[**Perceived Belief Distribution**] In the second part of this survey, you will again read statements on the same 10 political and socioeconomic issues. After reading each statement, you will be asked to provide your guesses about what percentage of other study participants agree or disagree with the statement, as well as how certain you are about your guesses. Note that all study participants are current students at UC Berkeley.

At the end of the study, we will randomly select one of your guesses. For this statement, you will have a chance to receive a bonus of $4 and you will maximize your chance of earning the bonus if you report your beliefs as accurately as possible. That is, there is nothing to gain by stating a number that differs from what you actually belief. You can find details about the bonus payment here [link to the Binary Scoring Rule].

(For each statement, participants get the following question): Among all Berkeley students who participate in this study, what percentage do you think will privately answer "agree" and "disagree" with this statement respectively? (please enter integers) Agree ___%; Disagree ___%

(Following their guesses, participants are also asked) How certain do you feel about your guesses above? (Choose between Certain and Uncertain)

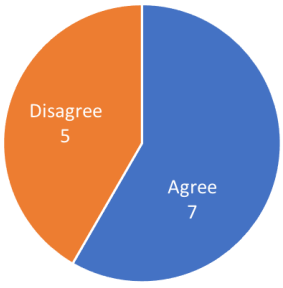
F.2 Midline Survey

The midline survey is sent 12 hours before participants' scheduled Zoom sessions.

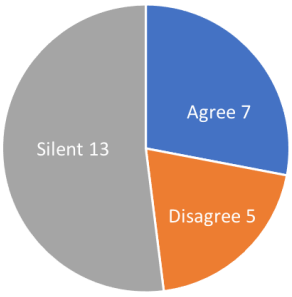
[**Welcome**] Thank you for completing the first part of our study. To help you get a better sense about the Zoom discussion, below we will show you summaries about earlier Zoom sessions conducted with other study participants who are also Berkeley students.

Then we present summary statistics on 4 topics that were discussed in Zoom sessions. We use one of the topics as an example here:

25 randomly selected Berkeley students like you discussed over Zoom if they agreed with the following statement: "If Proposition 209 was repealed, universities in the UC system should adopt extensive affirmative action policies that explicitly take into account race in the admission process." Here is a summary of their Zoom discussion:



Shown to the Control group



Shown to the Treatment group

- Among all Berkeley students who participate in this study, what percentage do you think will privately answer "agree" and "disagree"? (please enter integers)

  Agree ___%; Disagree ___%

- How certain do you feel about your guesses above? (Certain or Uncertain)

F.3 Endline Survey

The endline survey is distributed at the end of the Zoom session.

[**Perceived belief distribution**] *For each of the 4 topics, we again ask participants to provide their guesses about what percentage of other study participants agree or disagree with the statement.*

- Among all Berkeley students who partcipate in this study, what percentage do you think will privately answer "agree" and "disagree" with the following statement? (please enter integers) Agree ___%; Disagree ___%
- How certain do you feel about your guesses above? (Certain or Uncertain)

[**Expressions on a different topic**] In an earlier survey, we asked if you agreed or disagreed with the following statement. "Transgender athletes should not be allowed to compete in female sports at the college level."

We want to let Berkeley students know what other students think about this topic. This is your opportunity to tell other students what you think. If you would like to share your opinion with the other Berkeley students participating in this study, please indicate whether you agree or disagree with this statement and why. Your responses will be shared publicly with other study participants. You can also skip this question if you do not want to answer it or if you do not want to share your answers with other study participants.

[**Recall**] Now please take a moment to recollect the Zoom session that you just attended and answer the following question.

- How many students (excluding study moderators) were in the Zoom session?
- (For each statement) In the Zoom session, how many students agreed, disagreed or stayed silent when discussing the following statement? (please enter integers)
- (For each statement) Among those who stayed silent on this topic during the Zoom discussion, how many do you think privately "agree" and "disagree" respectively? (please enter integers)

F.4 Binary Scoring Rule

Determining your bonus: Your bonus will be determined by randomly selecting one of the statements to count and computing your payment according to the procedure below for the statement that counts.

1. You will state your guesses about the percentage of Berkeley students who privately agree/disagree with the statement. Denote your guess G (between 0 – 100) as the percentage of Berkeley students who privately answer "Agree".

2. The computer will randomly draw two numbers, X and Y, each with values between 0 and 100. For each draw, each number is equally likely to be selected. Draws are independent in the sense that the value selected for X in no way affects the value selected for Y and vice versa.

3. The computer will then randomly select one Berkeley student who participated in this study and get her anonymized answer (the answer will be kept private).

4. If the students' answer is "Agree", then you receive the bonus if your guess G is greater than or equal to either X or Y.

5. If the students' answer is "Disagree", then you receive the bonus if your guess G is smaller than either X or Y.

The procedure is designed so that you have the best chance of winning the bonus when you state your beliefs as accurately as possible about the fraction of Berkeley students you think privately agree or disagree with the statement.

F.5 Social Appropriateness of Statements

Following Krupka and Weber (2013), we elicit the social appropriateness of agreeing with each of the topic.

In this part of the survey, you will read statements on 10 political and socioeconomic issues. After reading each statement, you will be asked to evaluate whether the statement is "socially appropriate" and "consistent with moral or proper social behavior" or "socially inappropriate" and "inconsistent with moral or proper social behavior" at UC Berkeley. By socially appropriate, we mean behavior that most students at UC Berkeley agree is the "correct" or "ethical" thing.

At the end of the survey, we will randomly select one of the statements. For this statement, we will determine which response was selected by the most Berkeley students who answered

this survey. If you give the same response as that most frequently given by other students, you will receive an additional $2. For instance, if we randomly select the first statement and if your response had been "somewhat socially appropriate," then you would receive $2 if this was the response selected by most other students who answered this survey.

[**For each statement, we ask:**] Please indicate whether you believe holding this opinion is socially appropriate. (Choose between very socially inappropriate, somewhat socially in appropriate, neither socially appropriate nor inappropriate, somewhat socially appropriate, very socially appropriate.)

Appendix G Zoom sessions script

[Moderators: Have the following documents ready: 1) Slides; and 2) the random number sheet. Begin recording the Zoom session to the cloud as soon as the session begins.]

[**Introduction**] Thank you all for joining us today. This Zoom session is the second part of your participation in the study and should take approximately 50 minutes. My name is [moderator's name] and I will be moderating today's Zoom session. While we are waiting for the session to begin, please take 2 minutes to complete the brief survey we emailed you yesterday if you have not already done so.

To encourage active participation, we request that you keep your camera on throughout the Zoom session and your microphone muted unless you are called upon to speak. Also, if you haven't already done so, please take a minute to change your Zoom display name to your first name plus the first initial of your last name. Note that we will be recording this Zoom session for research purposes. Please keep everything that is discussed in this Zoom session private.

Before we get started, let's do a quick round of introductions. When I call your name, please unmute yourself and briefly tell us your year and major at Berkeley. And please let me know if I am pronouncing your name correctly. *[Conduct introductions, calling on participants in the order they appear on the moderator's screen. After each person you can say "thank you" or "welcome"]* Did I miss anyone?

Welcome everyone. To give you a quick overview of how we will structure our Zoom discussion, we have selected a few current socio-economic issues for us to discuss as a group. For each topic, I will first give you a brief introduction of the topic, then everyone who is interested will have a chance to express their views. Specifically, after introducing the topic I will read you a statement and you will have 90 seconds to decide if you'd like to share your views on the topic with everyone else in this Zoom room. If you would like to share your views, simply send a private chat message to me, the moderator, indicating whether you

"agree" or "disagree" with the statement. If you do not wish to share your views simply do not send a message to the moderator. I will then call on everyone who replied to one-by-one briefly state why they agree or disagree with the statement. Everyone who responds in the chat will be called upon, and I will randomly choose the order in which people are called upon. Please try to keep your comments respectful and under 1 minute. Please note that your study compensation will NOT depend on whether or not you share your views, or which views you share in the discussion. Your compensation just depends on whether you complete all parts of the study. Do you have any questions or comments before we begin?

[**Topic 1: Daylight Saving Time**] On March 2nd, a bill to make daylight saving time permanent in the U.S. was reintroduced in the Senate. The new Sunshine Protection Act is similar to the bill introduced last year. If passed, the clocks would not change again in November, or ever again. *[Moderator: Put in the chat: "Topic 1"]*

Now, we would like to hear your thoughts on daylight saving time. Do you agree or disagree with the following statement: "Daylight saving time should be eliminated". Please send the moderator a direct chat message saying "Agree" or "Disagree" if you would like to express your opinions. *(Discussion: Using the random order spreadsheet, call on everyone who responded via chat in a random order. You can respond "thanks for sharing" or "thank you" after each person. You can also say "30 seconds left" or "10 seconds left" if someone is going over time) [Conclusion for this topic]* Thank you everyone for sharing your views on this topic. Now I'll introduce the next topic.

[**Topic 2: Renaming Schools**] Since 2020, 82 schools across the country have been renamed because they honored controversial historical figures. In the Bay area, the San Francisco Unified School District voted on renaming 44 schools named after controversial public figures. *[Moderator: Put in the chat: "Topic 2"]*

Now we would like to hear your thoughts on this topic. Do you agree or disagree with the following statement: "All public schools named after controversial historical figures, including former Presidents George Washington, Thomas Jefferson, and Abraham Lincoln, should be renamed." Again, please send the moderator a direct chat message saying "Agree" or "Disagree" if you would like to express your opinions. *(Discussion: Using the random order spreadsheet, call on everyone who responded via chat in a random order. You can respond "thanks for sharing" or "thank you" after each person. You can also say "30 seconds left" or "10 seconds left" if someone is going over time) [Conclusion for this topic]* Thank you everyone for sharing your views on this topic. Now I'll introduce the next topic.

[**Topic 3: Death penalty**] The debate over the death penalty in the United States existed as early as the colonial period. Gallup has monitored support for the death penalty in the United States since 1937 in their surveys and observed gradual changes in respondents'

beliefs. We would like to understand what current college students think about the death penalty. *[Moderator: Put in the chat: "Topic 3"]*

Do you agree or disagree with the following statement: "The U.S. should abolish the death penalty." Again, please send the moderator a direct chat message saying "Agree" or "Disagree" if you would like to express your opinions. (Discussion: Using the random order spreadsheet, call on everyone who responded via chat in a random order. You can respond "thanks for sharing" or "thank you" after each person. You can also say "30 seconds left" or "10 seconds left" if someone is going over time) [Conclusion for this topic] Thank you everyone for sharing your views, now let's move on to the next topic.

[**Topic 4: Affirmative Action**] The Supreme Court is revisiting the use of ethnicity in college admissions policies, marking the third time in the last decade. The cases spotlight perennial questions about how highly competitive universities select incoming classes from a flood of applications and how they treat applicants of different racial backgrounds when reviewing their files. *[Moderator: Put in the chat: "Topic 4"]*

Do you agree or disagree with the following statement: "If Proposition 209 was repealed, universities in the UC system should adopt extensive affirmative action policies that explicitly take into account race in the admission process." Again, please send the moderator a direct chat message saying "Agree" or "Disagree" if you would like to express your opinions. *(Discussion: Using the random order spreadsheet, call on everyone who responded via chat in a random order. You can respond "thanks for sharing" or "thank you" after each person. You can also say "30 seconds left" or "10 seconds left" if someone is going over time) [Conclusion for this topic]* Thank you everyone for sharing your views, this was our final topic for today's session.

[Moderator: ONLY show the slides and use the script below if the previous discussion has been less than 35 minutes. Put in the chat: "Topic 5"]

[**Topic 5: Immunizations**] As schools enter the sixth semester of the pandemic, prevention measures such as vaccine mandates have become increasingly diverse and inconsistent. Schools across the country have adopted a wide range of vaccine policies based on CDC and public health guidance, state laws, religious beliefs and other factors.

Now, we would like to hear your thoughts on immunizations on Berkeley's campus. Do you agree or disagree with the following statement: "Immunizations, such as for Covid and the flu, should be required on Berkeley's campus." Again, please send the moderator a direct chat message saying "Agree" or "Disagree" if you would like to express your opinions. *(Discussion: Using the random order spreadsheet, call on everyone who responded via chat in a random order. You can respond "thanks for sharing" or "thank you" after each person. You can also say "30 seconds left" or "10 seconds left" if someone is going over time) [Conclusion for this*

topic] Thank you everyone for sharing your views, this was our final topic for today's session.

[**Conclusion**] That concludes our discussion today. Thank you all for participating, for being willing to share your views and listen to the views of others. To complete your participation in this study, please answer a final short survey. I just posted the link to the survey in the chat and it should take no more than 5 minutes to complete. You do have to complete this survey to meet all the requirements of this study and receive compensation. Feel free to turn off your video while you answer the survey or you may leave the Zoom room entirely if you prefer, please just complete it within the next 10 minutes. In the meantime, please let us know if you have any questions and thanks again for your participation today.

[Moderator: Keep the Zoom room open until the end of the scheduled slot or until the last participant has left. Before closing the Zoom room, please export the chat]