Breaking the Spiral of Silence *

Yuen Ho †

Yihong Huang ‡

February 9, 2024

Click Here for the Latest Version

Abstract

The Spiral of Silence theory plays a crucial role in contemporary political discourse. According to this idea, 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. We experimentally test this hypothesis based on a simple model in which self-censorship and limited attention to silence interact to jointly establish equilibrium norms. In our experiment, UC Berkeley undergraduates discuss controversial political and socioeconomic issues. Students with socially inappropriate views self-censor to a significant degree. Given the limited attention students pay to silence, self-censorship amplifies over time. We experimentally increase the salience of silence, and show that this affects both beliefs about others' views and public expression decisions. Because inference and expression amplify each other, different levels of attention to silence can produce divergent perceived social norms in equilibrium.

*We thank Benjamin Enke, Matthew Rabin, Gautam Rao, and David Yang for help and advice throughout different stages of this project. We would also like to thank Stefano DellaVigna, David Laibson, Jesse Shapiro, Edward Glaeser, Leonardo Bursztyn, Vincent Pons, Hunt Allcott, Jeremy Magruder, and participants at the Harvard Behavioral Economics Workshop, Harvard Political Economy Workshop, and UC Berkeley Psychology and Economics Workshop for valuable comments and discussions. We thank Madison Butolph 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). The experiment was pre-registered in the AEA registry, number AEARCTR-0009505. Financial support from the Kenneth C. Griffin Economics Research Funds and UC Berkeley Xlab Research Grants is gratefully acknowledged. All errors are our own.

†University of California at Berkeley, Department of Agricultural and Resource Economics. Email: yuen_ho@berkeley.edu

‡Harvard University, Economics Department. Job Market Paper and Corresponding Author, Email: yihong_huang@g.harvard.edu.

1 Introduction

Political discourse often features a silent majority overshadowed by a vocal minority. In 1969, Richard Nixon introduced the term "silent majority" to describe Americans not involved in the Vietnam war protests or the counterculture movement. Donald Trump later invoked the same term for his supporters during his presidential campaigns. 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 an environment where individuals self-censor due to potential social backlash, average expressed opinion will not accurately represent the actual distribution of views. In today's political climate, characterized by polarized views and an alleged "cancel culture", self-censorship may be especially pronounced. In particular, there is much discussion about self-censorship becoming more prevalent on college campuses, where strong norms around political correctness 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).

Self-censorship is typically viewed as problematic because the views of silent people are not represented in the decision making process. In a dynamic context, this can snowball into 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. When people pay limited attention to silence, they treat the distribution of expressed views as representative, which 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 present.

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? Specifically, we study the effect of increasing attention to silence on beliefs about others, public expression, and the resulting equilibrium of perceived social norms.

To formalize the spiral of silence and motivate our experiment, we develop a simple conceptual framework where there are two types of individuals, holding socially appropriate or inappropriate views. Individuals choose whether to express their views by weighing the intrinsic value of honest expression against potential social backlash. We assume

that the social sanction costs for expressing socially inappropriate views depend on the perceived dominance of socially appropriate views within society, while expressing appropriate views incurs no such costs. Thus, individuals with inappropriate views, expecting harsher sanctions, are more likely to self-censor.[1] Simultaneously, individuals observe the publicly expressed views and silence, using these signals to update their beliefs about others' views. The updated beliefs about others, in turn, inform their future expression decisions. Assuming a Bayesian interpretation of silence, individuals who pay more attention to silence have a lower estimate about the popularity of socially appropriate views, which increases their willingness to express inappropriate views. We further derive the equilibrium of perceived social norm and public expression decisions. Notably, the perceived social norm in equilibrium is closely tied to the degree of attention to silence. In equilibrium, in a society of rational agents fully attentive to silence, the perceived opinion distribution converges to the actual distribution, while in a society with limited attention to silence, people overestimate the prevalence of socially appropriate views.

Given recent discussions about self-censorship on liberal college campuses, we study these questions using an experiment with UC Berkeley undergraduate students. We test the three central elements of the spiral of silence theory: 1) the prevalence of self-censoring; 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 interact in small groups over Zoom to discuss controversial political and socioeconomic topics. Before their own group discussions, students are presented with summary statistics from previous groups, wherein we randomly vary the salience of silence. The topics discussed included: 1) whether all public schools named after 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.

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 the group discussion without any additional information. Among other things, the First Movers sample allows us to measure how private views

[1] We further assume that the psychological cost of misrepresenting one's authentic beliefs is so substantial that, where possible, individuals prefer silence over lying.

predict self-censorship in group discussions. We find that self-censorship occurs among college students and is meaningful in magnitude. First Movers holding socially appropriate views are approximately 14% more likely to express their opinions publicly in Zoom sessions compared to those who hold socially inappropriate views.

Our experimental setup is dynamic, intended to capture how students' perceptions about social norms evolve after observing the stated opinions of others. Specifically, we share a summary of the views expressed publicly by First Mover with the remaining study participants (Second Movers). To causally identify the effect of attention to silence, we randomly vary whether participants receive "Control" information, which shares the number of First Movers who publicly agreed or disagreed with each topic, or "Treatment" information, which includes the Control information plus information on the number of First Movers who remained silent. For example, Control participants see a pie chart showing that 7 First Movers publicly agreed with renaming schools and 5 First Movers 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 calculate the number of silent participants from the information provided. The key difference between the Control and Treatment conditions is 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. We then invite participants to Zoom sessions of either all Control or all Treatment participants to discuss the same topics with other students. 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 the dynamics in societies where people pay similar level of attention to silence.

Our results indicate that increasing the salience of silence affects both inference and expression. Drawing attention to silence reduces the perceived popularity of the socially appropriate view. Treatment students guess that the socially appropriate view is on average 6% less prevalent compared to those in the Control students.

Increasing attention to silence also affects public expression. Participants in the Treatment group who privately hold the socially inappropriate view are about 17% more likely to publicly express their opinions in Zoom sessions relative to the Control group. Moreover, effects on expression appear to be driven by changes in perceptions about public opinion. When we use treatment assignment as an instrument for beliefs, a 1 percentage point decrease in the perceived popularity of socially appropriate views is associated

with a 2.5% higher probability of expressing socially inappropriate views. As predicted by theory, there are no significant treatment effects on expression for those who hold the socially appropriate view.

Moreover, the effects on inference and expression appear to amplify and build on each other, possibly resulting in divergent equilibria. At baseline, the views of treatment and control individuals are comparable. After the information intervention, treatment participants guess that socially appropriate views are on average 6% less prevalent compared to control participants. After participating in the Zoom sessions, the difference in beliefs between treatment and control participants increases to 10%. The effects are consistent with the information treatment initially changing the perceived belief distribution and therefore the expression of views in the Zoom session, which then further changes perceptions of public opinion among treatment and control participants.

Our experiment consists of only two rounds of expression and three rounds of belief-updating. Yet an important insight that emerges from the results is that the effects on inference and expression are dynamic and self-reinforcing, which calls for an equilibrium analysis. To characterize equilibria with different levels of attention to silence, we structurally estimate our model. The structural estimates suggest that our information treatment increases attention to silence by around 40 percentage points. Using the estimated parameters, We calculate two benchmarks for perceived social norms in equilibrium, with full attention to silence and no attention to silence. In the case where everyone in a society pays full attention to silence, the perceived belief distribution converges to the truth. Conversely, the case where everyone is completely inattentive to silence leads to an equilibrium where the popularity of the socially appropriate view is overestimated by 50 percentage points.

Taken together, our experiment generates a few key insights. First, the spiral of silence exists in practice and gives rise to an equilibrium in which only socially appropriate views are expressed and considered dominant. The Control condition mirrors the amount of attention individuals typically pay to silence when their attention is not drawn explicitly to it. When a topic is socially sensitive, those who hold the socially inappropriate view are more likely to stay silent. When attention paid to silence is limited, this initial silence by people who hold socially inappropriate views distorts beliefs about the prevailing norm. This, in turn, reinforces self-censorship and exacerbates the belief distortion.

Second, the perceived social norms and views expressed in equilibrium depend on the level of attention that people pay to silence. Drawing attention to silence has significant impacts on both inference and expression. Treatment participants pay more attention to silence and thus more accurately infer about the true popularity of the socially appropriate

view when observing the stated views of others. This, in turn, makes them more willing to express socially inappropriate views publicly, breaking the spiral of silence.

This latter finding has important policy implications for how to design and share information from commonly used public feedback processes, such as opinion polls, public petitions, or town halls. Such feedback processes typically only report on 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.

To test the external validity of our results, we conduct an extension of our main experiment and find that increasing attention to silence not only affects political discourse but also political action, namely the willingness to sign a petition. To do so, we focus on a real public feedback process conducted by the Berkeley Building Name Review Committee. As part of its decision process, the Committee asks for public comments for each proposed renaming. Community members can share their comments publicly on the Committee's website, or keep their comments confidential and visible only to the Committee. Following the same design, we randomly assign participants to either a Control group or a Treatment group where they see different pie charts summarizing these comments excluding or including the number of confidential comments respectively. Mirroring findings from our main experiment, students in the Treatment group, who saw confidential comments counts, estimated fewer peer students privately supporting renaming. Extending our main results, we further find that those in the Treatment group who privately disagree with renaming are 20% more likely to sign a public petition against the renaming of other Berkeley buildings, relative to the Control group.

Beyond college campuses, we test a distinctive prediction of our model about the correlation between silence and misperceptions in nationally representative surveys. Under the assumptions of our stylized model, 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 could stifle public expression, and silence in turn can lead to less accurate beliefs about others. We test for this positive correlation with data from American National Election Studies (ANES), a series of nationally representative surveys on public opinion and political participation. ANES surveys are conducted mostly face-to-face or over phone, and include questions asking respondents to position themselves, Democratic, and Republican parties on a 7-

point scale for various political and socioeconomic issues over the past five decades, since 1970. 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. We indeed find a positive correlation between the share of non-responses and the magnitude of misperceptions. After controlling for year, topic, party fixed effects, and actual beliefs, a 10% increase in the share of silence is correlated with a 2.6 percentage point increase in misperceptions about others views on a topic.

Overall, the main contribution of our paper is threefold. First, we document how self-censorship and limited attention to silence lead to a spiral of silence, and that this generates an equilibrium where people overestimate the prevalence of socially appropriate views. Second, we document that the level of attention to silence determines, at least in part, beliefs and expression in equilibrium. This is of direct policy relevance, especially on how to design and share information from public feedback processes. Third, we provide suggestive evidence that silence plays a critical role not only on college campuses, but also in political discourse in general.

Related Literature. This paper is closely related to the political science literature on the Spiral of Silence, introduced in Noelle-Neumann (1974).[2] Existing studies on the spiral of silence focuses 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). However, one crucial perspective that is rarely studied in this literature is how people interpret silence. In this paper, we present direct evidence on the role of silence in the formation and breakage of the spiral of silence. We also add to this literature by studying beliefs and expression in equilibrium.

Our study also contributes to a growing literature showing that misperceptions about others are widespread and that correcting such misperceptions can lead to meaningful changes in behaviors (Bursztyn, Egorov, and Fiorin, 2020; Bursztyn, González, and Yanagizawa-Drott, 2020; Bursztyn and Yang, 2021).[3] One paper that also studies political correctness and misperceptions on college campuses is Braghieri (2021). Different

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

[3]There exists multiple conceptual frameworks to explain the persistence of misperceived social norms, including 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.

from Braghieri (2021), our experiment specifically studies the role of silence, and shows how self-censorship and inattention to silence jointly produce equilibrium norms. We contribute to this literature by providing evidence that inattention to silence can explain the origin and persistence of misperceived social norms.

This paper also builds on existing evidence that people do not correctly learn from "nothing": see lab evidence in Esponda and Vespa (2018); Enke (2020); Jin, Luca, and Martin (2021), and applications in finance and marketing in Hirshleifer and Teoh (2003); Li and Hitt (2008); Koehler and Mercer (2009); Giglio and Shue (2014). In this paper, we apply this concept to the domain of political discourse, where silence and misperceptions are widespread and have economically meaningful impact.[4]

In what follows, we present a motivating framework in Section 2. In Section 3, we describe the experimental design, sampling, and timeline. In Section 4, we present the main experimental results on inference, expression, and dynamics. We structurally estimate the model and calculate beliefs in equilibrium in Section 5. In Section 6, we discuss attention to silence as a key mechanism that drives the experimental results. We discuss the external validity of the experiment with two extensions in Section 7. In Section 8, we discuss lessons from the study and implications for policy.

2 Motivating Framework

To motivate the experimental design, we present a simple model where an individual's decision to express an opinion publicly depends both on their true private views and their perceptions of others' opinions. Importantly, these perceptions are shaped by both the views that others choose to express publicly and the choice of others to stay silent.

Suppose that there is a continuum of individuals indexed by i, with heterogeneous private preferences θ_i over some policy, where $\theta_i \in \{A, D\}$. For example, $\theta_i = A$ if individual i agrees (A) with the policy and $\theta_i = D$ if i disagrees (D) with the policy. We assume that $\theta_i = A$ (agree) is the socially appropriate view.

[4]Finally, this paper speaks to a large behavioral economics literature on limited attention, especially bottom-up attention (Bordalo, Gennaioli, and Shleifer, 2012, 2013, 2022). Empirical evidence shows that shifting attention to nonsalient or opaque features have significant impact on economic decisions such as consumption and production (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) In our context, expressed signals are often more salient stimuli than unexpressed signals, attracting attention. Our study exogenously varies attention to silence and demonstrates the subsequent impacts on both inference and public expression.

2.1 Expression

When considering whether to express their opinions publicly, individuals both derive intrinsic value from voicing their opinions honestly $V_i \sim N(v, \sigma_v^2)$, and also disutility from potential social backlash if they express an opinion that violates the social norm. Therefore, for someone who disagrees with the policy ($\theta_i = D$), her utility of expressing the opposition view is:

$$U_i^D(e_i = D) = V_i - \chi \cdot E[\pi_i]$$

where $e_i \in \{A, 0, D\}$, $e_i = D$ if she honestly expresses her opinion and $e_i = 0$ if she stays silent. We assume that the psychological cost of misrepresenting one's authentic beliefs (e.g. $e_i = A|\theta_i = D$) is so substantial that, where possible, individuals prefer silence over lying. Detailed discussion about distorted public expression can be found in Section 2.4. $\chi > 0$ denotes the strength of social sanction; and π_i is individual i's prior about the fraction of people who hold the socially appropriate view. Suppose that the prior about the belief distribution for each individual i in the society is homogeneous, given by $\pi_i \overset{\text{iid}}{\sim} Beta(a, d)$, $E[\pi] \equiv E_i[E[\pi_i]] = E[E_i[\pi_i]] = \frac{a}{a+d}$. [5]

Intuitively, individuals trade-off between honestly speaking up and avoiding potential social backlash, which depends on both how "inappropriate" their view is (captured by χ) and how "unpopular" their view is (captured by $E[\pi]$). For simplicity, we assume that $U_i^D(e_i = 0) = 0$. Therefore, among individuals with $\theta_i = D$, those with $V_i \geq \chi E[\pi]$ will speak up. In society, the fraction of those who honestly express D is

$$Pr(e_i = D|\theta_i = D) = 1 - G(\chi E[\pi]) \tag{1}$$

where G is the cdf of V_i. Likewise, those who hold the socially acceptable view do not experience any social backlash when expressing their opinions, resulting in a fraction of those who honestly express A as

$$Pr(e_i = A|\theta_i = A) = 1 - G(0) \tag{2}$$

It directly follows that those who hold the socially inappropriate view (D) are less likely to express their opinions than those who hold the socially appropriate view (A) because of potential social backlash $G(\chi E[\pi]) > G(0)$.

In a society with size N, when following the decision rule above, the observed number

[5] Discussions about heterogeneous priors can be found in Appendix A.3

of people who express A, D, or who stay silent are, respectively,

$$S_A = [1 - G(0)]pN$$
$$S_D = [1 - G(\chi E[\pi])](1-p)N$$
$$S_S = [G(0)p + G(\chi E[\pi])(1-p)]N$$

where $p = Pr(\theta_i = A)$ is the actual fraction of individuals with $\theta_i = A$. Denote $\mathcal{S} = (S_A, S_D, S_S)$ as the set of expressed and unexpressed signals.

2.2 Inference

We assume that the prior about the fraction of people who hold the socially appropriate view for each individual i in society is $\pi_i \overset{\text{iid}}{\sim} Beta(a, d)$, $E[\pi] \equiv E_i[E[\pi_i]] = \frac{a}{a+d}$. We make the assumption that π_i is i.i.d because the concept of "social norms" is closely tied to collective and shared expectations (Gibbs, 1965; Lapinski and Rimal, 2006). We relax this assumption in Appendix A.3. People observe the publicly expressed and unexpressed opinions $\mathcal{S} = (S_A, S_D, S_S)$ and update their beliefs according to these signals. For simplicity, we assume Bayesian updating is applied to signals that are attended to. We discuss several other potential updating rules in Appendix A.4.

For individuals who are fully inattentive to silence, their belief-updating process depends solely on the expressed opinions and their posteriors, denoted by $\gamma_0 \sim Beta(a + S_A, d + S_D)$, and

$$E[\gamma_0|\mathcal{S}] = \frac{a + S_A}{a + d + S_A + S_D} \tag{3}$$

For individuals who pay full attention to silence and infer from silence in a Bayesian way, they infer that

$$Pr(\theta_i = k|e_i = 0) = \frac{Pr(e_i = 0|\theta_i = k)Pr(\theta_i = k)}{\sum_k Pr(e_i = 0|\theta_i = k)Pr(\theta_i = k)}, k \in \{A, D\}$$

Their posteriors are thus given by

$$\gamma_1|\mathcal{S} \sim Beta(a + S_A + \frac{aG(0)S_S}{aG(0) + dG(\chi E[\pi])}, d + S_D + \frac{dG(\chi E[\pi])S_S}{aG(0) + dG(\chi E[\pi])})$$

$$E[\gamma_1|\mathcal{S}] = \frac{a + S_A + S_S \cdot \frac{a}{a + d \cdot G(\chi E[\pi])/G(0)}}{a + d + S_A + S_D + S_S} \tag{4}$$

More generally, we can model partial attention to silence and use $\lambda \in [0, 1]$ to denote the level of attention paid to silent individuals.

$$E[\gamma(\lambda)|\mathcal{S}] = \frac{a + S_A + \lambda S_S \cdot \frac{a}{a+d \cdot G(\chi E[\pi])/G(0)}}{a + d + S_A + S_D + \lambda S_S} \tag{5}$$

The case where $\lambda = 0$ gives us Equation 3, in which people are completely inattentive to silence, and the case where $\lambda = 1$ gives us Equation 4, in which people pay full attention to silence. Examining the comparative statics with respect to λ leads to our first proposition:

Proposition 1. *Assuming uninformative priors, if the socially appropriate view is expressed more often ($S_A \geq S_D$), $E[\gamma(\lambda)|\mathcal{S}]$ decreases in λ.*[6]

Proof: With uninformative priors,

$$\frac{\partial E[\gamma(\lambda)|\mathcal{S}]}{\partial \lambda} \propto a[1 - \frac{G(\chi E[\pi])}{G(0)}] + S_D - S_A \frac{G(\chi E[\pi])}{G(0)} < 0 \tag{6}$$

In other words, individuals who pay more attention to silence have a lower expected posterior belief about the true popularity of the socially appropriate view. In the experiment, we test Proposition 1 by exogenously varying whether participants' attention is drawn to those who stayed silent in a public discussion, hence experimentally increasing λ.

2.3 Equilibrium

As noted by Matthes (2014): "Time is of the utmost importance when designing studies to test the spiral of silence theory... Few studies have tested the dynamic nature of the theory." Our model tries to capture the dynamics within the intertwined expression and belief-updating processes. In each round t, individuals make expression decisions based on the perceived prevalence of the socially appropriate view $E[\pi_t]$. These decisions determine the views expressed (signals) in the following round: $S_{A,t}$, $S_{D,t}$ and $S_{S,t}$, from which individuals further update their beliefs $E[\pi_{t+1}]$.

$$E[\pi_{t+1}(\lambda)|S_t] = \frac{a_t + S_{A,t} + \lambda S_{S,t} \cdot \frac{a_t}{a_t + d_t \cdot G(\chi E[\pi_t])/G(0)}}{a_t + d_t + S_{A,t} + S_{D,t} + \lambda S_{S,t}} \tag{7}$$

In equilibrium, with full attention to silence and rational updating, the equilibrium

[6]Some common choices of uninformative priors include: uniform (Bayes-Laplace) prior ($a = d = 1$), Jeffreys prior ($a = d = 1/2$), Haldane prior ($a = d = 0$) or its approximation ($a = d = \varepsilon$)

beliefs $\pi^* \sim (a^*, d^*)$ and expressions S_A^*, S_D^*, S_S^* satisfy the following equations:

$$\begin{aligned} E[\pi^*(\lambda=1)] = \frac{a^*}{a^*+d^*} &= \frac{S_A^* + S_S^* \cdot \frac{a^*}{a^*+d^* \cdot G(\chi E[\pi^*(\lambda=1)])/G(0)}}{S_A^* + S_D^* + S_S^*} \\ S_A^* &= [1-G(0)]p \\ S_D^* &= [1-G(\chi E[\pi^*(\lambda=1)])](1-p) \\ S_S^* &= G(0)p + G(\chi E[\pi^*(\lambda=1)])(1-p) \end{aligned} \tag{8}$$

where p is the actual fraction of individuals who hold the socially appropriate view: $\theta_i = A$. In contrast, if people are completely inattentive to silence, the equilibrium beliefs should satisfy the following condition:

$$E[\pi^*(\lambda=0)] = \frac{a^*}{a^*+d^*} = \frac{S_A^*}{S_A^*+S_D^*} = \frac{[1-G(0)]p}{[1-G(0)]p + [1-G(\chi E[\pi^*(\lambda=0)])](1-p)} \tag{9}$$

Proposition 2. *With Bayesian updating (i) for agents who pay full attention to silence : $E[\pi^*(\lambda=1)] = p$. (ii) for agents who pay no attention to silence: $E[\pi^*(\lambda=0)] > p$. (iii) $E[\pi^*(\lambda)]$ decreases in λ, where $\lambda \in [0,1]$,*

Proof of Proposition 2 is detailed in Appendix A.1. Intuitively, rational agents pay full attention to silence and understand the selection bias into silence. In equilibrium, their 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. Our model also predicts that agents who pay no attention to silence will overestimate the popularity of the socially appropriate view in equilibrium. This is because such agents only update their beliefs according to the publicly expressed opinions, which are skewed in the direction of the socially acceptable view.

Simulations illustrate divergent paths to equilibrium beliefs, as shown in Figure 1. It depicts the evolution of beliefs for 1000 individuals over 50 periods, where A (*Agree*) is considered socially appropriate, yet only privately supported by 45% of the population. The assumption of homogeneous priors is relaxed. The red line represents the belief trajectory with complete inattention to silence ($\lambda = 0$), revealing a spiral of silence where the socially appropriate view A is perceived to be increasingly dominant. The blue line, with full attention to silence ($\lambda = 1$), gradually converges to the actual belief distribution. Meanwhile, the green line demonstrates a scenario with partial attention to silence.

Experimentally, we test whether varying attention to silence can explain differences in equilibrium beliefs by grouping participants into Zoom sessions based on their treatment

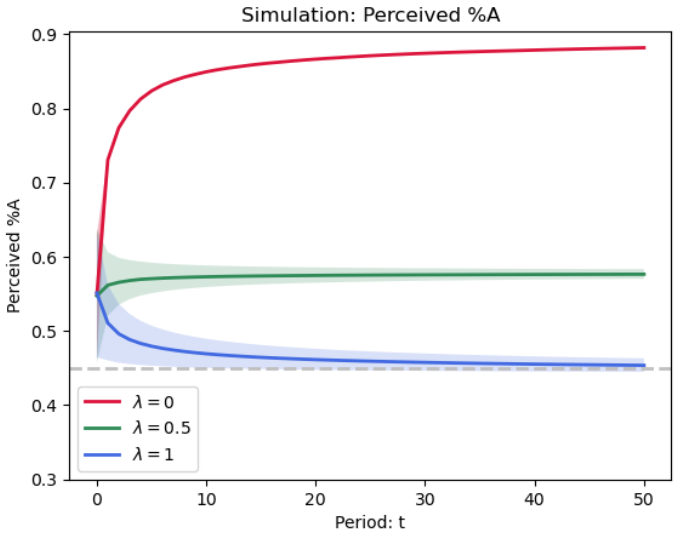


Figure 1: Simulation, Perceived Fraction with $\theta = A$

Note: This figure shows the simulation with 1000 people over 50 periods. A is the socially appropriate view. The actual fraction of people in favor of A is 45%. In each period, individuals make expression decisions, observe the expressed signals and update their beliefs about $\%A$. Individual i's prior about $\%A$ follows Beta distribution $Beta(a_i, d_i)$. To introduce heterogeneity in priors, we randomly draw a_i and d_i from $N(\bar{a}, \sigma^2)$ and $N(\bar{d}, \sigma^2)$, where $\bar{a} = 5.5, \bar{d} = 4.5, \sigma = 0.15$. V_i is drawn from the $N(0, 1)$ and the social sanction parameter χ is set to be 2.

assignment and measuring their beliefs about others' views at each stage.

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

Proposition 3. *Assuming $E[\pi] > p$ and $\lambda \in [0, 1)$, misperception ($\Delta \equiv E[\pi] - p$) is positively correlated with share of silence: $Corr(S_{S,t}, \Delta_t) > 0$, $Corr(S_{S,t}, \Delta_{t+1} | \mathcal{S}_t]) > 0$.*

Proof of Proposition 3 can be found in Appendix A.2. 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 3 with the ANES data.

2.4 Discussion

Neither lying nor changing private opinions are encompassed in our model. Under our framework, for those who hold the socially inappropriate view, the act of lying introduces an extra psychological discomfort in comparison to staying silent. Therefore, we operate under the presumption that staying silent is generally preferable to lying. Indeed, our empirical results suggest that lying is not a first order occurrence in our setting. In our

experiment, no more than 1% to 2% of participants voiced an opinion opposite of the one they reported privately holding in each Zoom discussion topic. This can either be interpreted as lying or as a change in opinion during the experiment.

We also assume that opinions (θ_i) 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. Our model can be generalized to accommodate multiple categories of private preferences. The main prediction holds that the perceived popularity of the socially appropriate category decreases in the level of attention paid to silence (λ). See Appendix A.5 for details. In this case, it is reasonable to assume that individuals holding more extreme views derive higher intrinsic utility from public expression. When this differential expression tendency (based on the strength of views) is combined with inattention to silence, perceptions about the distribution of views will be distorted. Specifically, extreme views will be perceived as being more widespread than they actually are, thereby increasing perceived polarization. In Appendix E, we show that experimentally drawing attention to silence reduces perceived polarization.[7]

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 whether socially appropriate views are indeed expressed more often publicly. To test whether inference from silence affects beliefs about others and public expression over time, we experimentally vary the salience of silence when sharing information about discussion 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 the spiral of silence theory.

[7]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.

3.1 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, for example Renaming Schools, Affirmative Action and Transgender Athletes. 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.[8] 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.[9] We also collect demographic information in the baseline survey, including gender, ethnicity, school year, major, and self-identified political affiliation.

3.1.2 Zoom Sessions

Zoom 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 similar 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 their first name and the first initial of their last name. Each Zoom session is facilitated by a trained moderator.

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. Specifically, participants first

[8]Participants can also choose to skip any questions if they do not want to provide their private views. Only 1% of participants skipped questions asking about their private beliefs in the baseline survey.

[9]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 G.

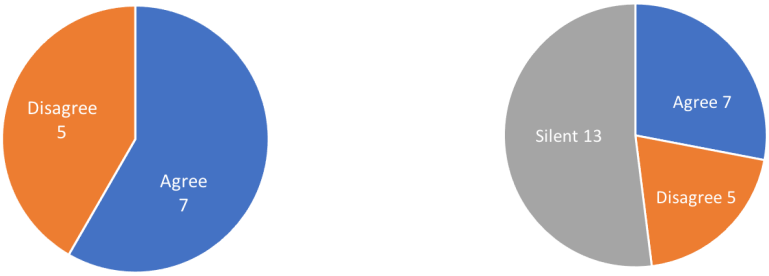
message the moderator privately indicating whether they "agree" or "disagree" with the statement if they want to share their views publicly. Students who do not wish to share their views simply do not message the moderator. The moderator then calls upon all of the participants who volunteered in a random 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 discussion, 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 H.

First Movers: A random subset of participants are 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 2). Treatment participants receive the same 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, we inform Treatment 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", 5 "Disagree" and 13 "Silent" (see the right panel of Figure 2). Note that both Control and Treatment groups are provided with sufficient information to determine the number of participants who stayed silent on each topic. The key difference between the Control and Treatment information is the *salience* of silence, which allows us to isolate the causal effects of increasing attention to silence on subsequent inference and expression decisions.

The information treatment is embedded into 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 a set of socially sensitive topics.

For the Second Movers Zoom sessions, Control participants are only eligible to sign



Control group Treatment group

Figure 2: Information Treatments

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. The protocol for the Zoom sessions for all Second Movers, including Control and Treatment participants, is otherwise identical to that for 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 lasted approximately 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.

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 the Zoom discussions they participated in. In particular, we ask participants to recall how many students publicly expressed a view that agreed or disagreed with the topic statement or stayed silent for each topic discussed in their Zoom session. 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 3 illustrates our experimental design in full and detailed survey instruments can be found in Appendix G.

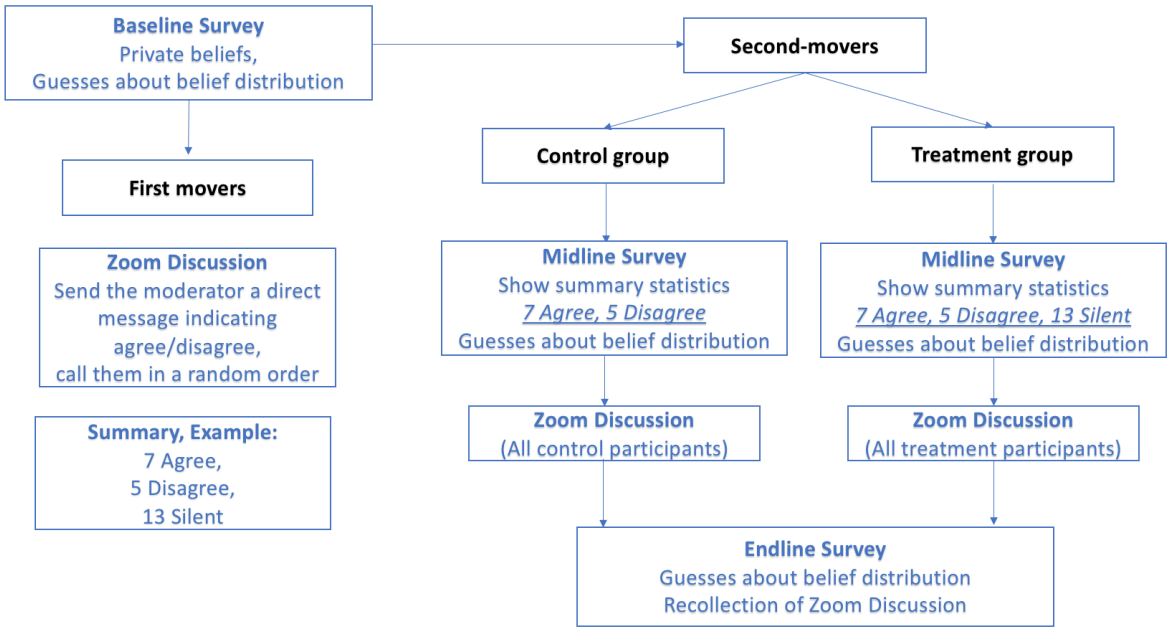


Figure 3: Experimental Design

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.[10] For our main experiment, we 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 Saving Time "DST" (Placebo): Daylight saving time should be eliminated.

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

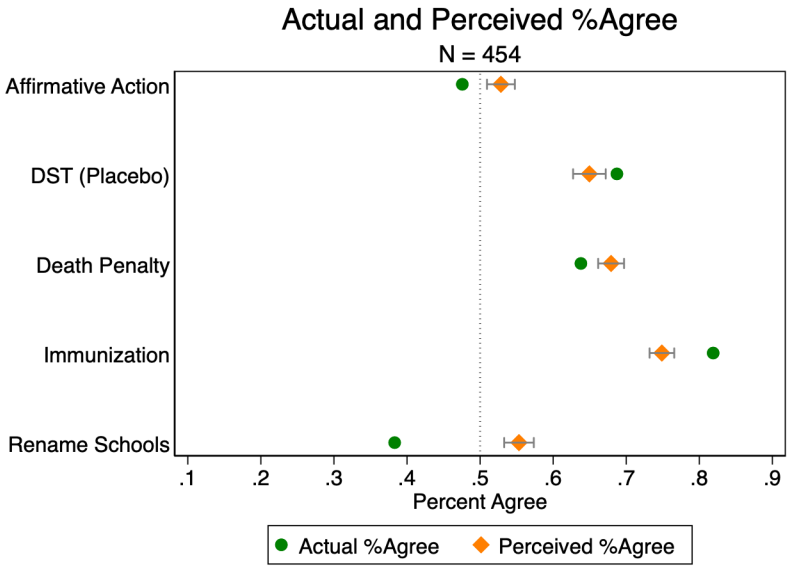


Figure 4: Baseline Actual and Perceived Fraction of Student Agreement

In the baseline survey, we elicited both the private beliefs and the beliefs of other students' views on each of these topics. Figure 4 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". From these results from the baseline survey, we observe that misperceptions are widespread: 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 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%.

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 were asked to rate whether each of the statement is "very appropriate", "somewhat inappropriate", "neither appropriate nor inappropriate", "somewhat appropriate", or "very appropriate", and they were incentivized to match their responses with the modal response. 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.

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.7 in Appendix B presents the percentage of students who chose

"very inappropriate", "somewhat inappropriate", "neither appropriate nor inappropriate", "somewhat appropriate", and "very appropriate" for each statement, with bold numbers representing the modal response for each topic.

We have standardized the direction of each statement so that "Agree" consistently corresponds to the socially appropriate view. For example, a majority of participants ("Somewhat Appropriate" (45.26%) + "Very Appropriate" (30.53%) = 75.79%) believe that it is socially appropriate to agree with the statement that the U.S. should abolish the death penalty. Nearly half of the participants consider the statement about eliminating daylight saving time to be neutral.

Apart from the five topics that are discussed in Zoom sessions, we also include some other political and socioeconomic issues in the baseline survey to reduce experimenter demand effect. A detailed description of these topics can be found in Appendix B.1. In Appendix B.2, we also discuss in detail why we select the five topics above to be included in Zoom 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 each round, we recruited approximately 100 participants. 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. We collected around 100 responses for the baseline survey each week and invited participants to Zoom sessions scheduled for the following week.[11] 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 other rounds. 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. The majority of participants identify themselves as "Liberal" or "Slightly liberal" (see Panel A in Table C.8 in Appendix C).

[11] Specifically, we recruited 95, 101, 63, 88 and 107 participants during the week of Feb 14, Feb 28, Mar 7, Mar 14 and Mar 28 respectively.

Among participants recruited in the first round, we randomly assigned 50 students 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.8).

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 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 Control/Treatment assignment, observable characteristics, and baseline views and beliefs (see Table C.9 in Appendix C).

3.4 Discussion: From Model to Experiment

Our experimental design allows us to answer three main research questions. First, by comparing the views expressed privately in the baseline survey versus publicly in Zoom sessions among First Movers, we test whether those who hold socially appropriate views are more likely to speak up. In the stylized model, we assumed that people with socially inappropriate views have social sanction costs to publicly dissent, 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 private views and public expression decisions.

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 see its impact on beliefs about others. This is a direct test of Proposition 1.

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 2 predicts that different levels of attention to silence produce divergent perceived social norms in equilibrium, through the self-reinforcing effect of self-censorship and limited attention to silence. Our experiment incorporates several rounds of inference and expression to shed lights on the equilibrium outcomes.

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. For analysis, we standardize responses so that the "Agree" stance is perceived as the socially appropriate view. In the following sections, we combine these four topics when presenting the main results. We also included Daylight Saving Time as a placebo (socially neutral) topic in the Zoom sessions, and results on this placebo topic are reported 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 unpopular or 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:

$$Express_{i,j} = \beta Agree_{i,j} + X_i + \alpha_j + \varepsilon_{i,j} \tag{10}$$

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 $\%Agree$. To account for potential correlations in the expression residual within subjects, we cluster standard errors at the individual level. As discussed in Section 3, all responses have been standardized so that $Agree$ corresponds to the socially appropriate view. 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 14% 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 10. 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.

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	0.142**	0.135*	0.138*	0.147*
	(0.0699)	(0.0728)	(0.0718)	(0.0784)
Panel B: Logit				
Private Agree	0.147**	0.140*	0.144**	0.133*
	(0.0706)	(0.0724)	(0.0707)	(0.0756)
Topic FE	✓	✓	✓	✓
Baseline guesses		✓	✓	✓
Session FE			✓	✓
Ind Controls				✓
Mean	0.470	0.470	0.470	0.470
SD	0.501	0.501	0.501	0.501
IDs	50	50	50	50
Obs	200	200	200	200

Standard errors clustered at individual level.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports the public expression decisions by First Movers on four topics discussed in Zoom sessions (Affirmative Action, Death Penalty, Immunizations, and Renaming Schools). $y_{i,j} = 1$ 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.

Recall that the primary objective of having First Movers in our experiment is to generate information on public expression which then informs the information treatments we provide to 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.10 in Appendix C displays the summary statistics from the selected First Mover Zoom sessions that were presented to the Control and Treatment groups.[12] Similar to Figure 2, we showed different pie charts excluding or including the number of First Movers who remained silent for the Control and Treatment groups respectively.

[12]Table C.11 report the summary statistics if we used data from all First Movers Zoom sessions. "%Agree", "%Disagree" and "%Silent" are similar as in Table C.10.

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 prevalence of socially appropriate views. Specifically, we estimate:

$$\%\tilde{Agree}^{M}_{i,j} = \beta_1 Treat_i + \beta_2 Agree_{i,j} + \beta_3 \%\tilde{Agree}^{B}_{i,j} + \alpha_j + X_i + \varepsilon_{i,j} \tag{11}$$

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 summary statistics; $Treat_i$ 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; 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 5). 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 11. On average, participants in the Treatment group guess that the socially appropriate view is 6.7% less prevalent than the guesses of 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.

Taken together, our results suggest that increasing attention to silence leads to lower estimates of the prevalence of socially appropriate views, as predicted by Proposition 1 in the conceptual framework. Furthermore, it appears that the Control information is not a neutral information treatment but rather moves the beliefs of Control participants towards the views publicly expressed by First Movers.

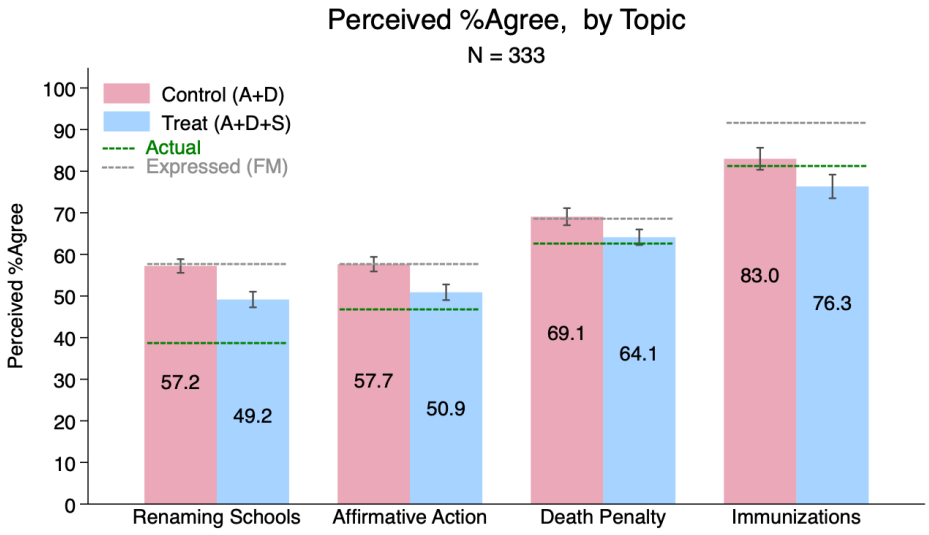


Figure 5: Perceived Popularity of the Socially Appropriate View at Midline

Note: This graph depicts the perceived share of agreement for each topic, as elicited in the midline survey. The white dashed line shows the actual share of agreement elicited in the private baseline survey. The green dashed line displays the expressed share of agreement that participants are shown as part of both the Control and Treatment interventions. 95% confidence intervals are shown in the graph.

Table 2: Perceived Popularity of the Socially Appropriate View 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***	-9.251***
	(0.970)	(0.930)	(0.958)	(1.490)	(1.481)	(1.176)
Private Agree		3.860***	3.198***		3.664***	2.693***
		(0.858)	(0.850)		(0.926)	(0.879)
$\%\tilde{Agree}^B$		0.115***	0.116***		0.110***	0.0855***
		(0.0241)	(0.0250)		(0.0270)	(0.0236)
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

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: The outcome variables are perceived $\%Agree$ elicited in the midline survey and the endline survey respectively. Combined results from the four topics (Affirmative Action, Death Penalty, Immunizations, and Renaming Schools) are reported in this table. 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.

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 separate public forums and then comparing differences in 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 all individuals who either pay attention to silence (Treatment) or do not (Control) will diverge to different perceived social norms and public political 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. Specifically, we test whether individuals who privately hold the socially inappropriate view are more likely to express those views publicly in the Treatment group relative to the Control group. To do so, we estimate the following Logit regression for individual i who privately disagree with statement j:

$$PublicDisagree_{i,j} = \beta_1 Treat_i + X_i + \theta_k + \varepsilon_{i,j} \tag{12}$$

We also estimate the following regression to measure the causal effects of increasing attention to silence on endline guesses about the belief distribution:

$$\%\tilde{Agree}^{E}_{i,j} = \beta_1 Treat_i + \beta_2 Agree_{i,j} + \beta_3 \%\tilde{Agree}^{B}_{i,j} + X_i + \theta_k + \varepsilon_{i,j} \tag{13}$$

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 a moderator fixed effect. Standard errors are clustered at the Zoom session level in both Equations 12 and 13.

Our results indicate that increasing attention to silence significantly increases the likelihood that an individual publicly expresses a socially inappropriate view. Specifically,

for our preferred specification which corresponds to Equation 12, as represented by column (5) in Table 3, participants in the Treatment group who privately disagree with the socially appropriate view are about 17% more likely to publicly express their opinions, holding other factors constant, relative to the Control group. As predicted, there are no significant treatment effects for those who privately agree with the socially appropriate view. These results are robust across different specifications, including using a OLS or Logit model and using different sets of control variables. Note that Table 3 pools results across all four Zoom topics to maximize power, results by topic can be found in Table C.12 in Appendix C.

Table 3: Public Expression Decisions by Second Movers

	OLS (Express = 1)				Logit (Express = 1)
	(1)	(2)	(3)	(4)	(5)
Panel A: Privately Disagree					
Treat	0.160***	0.161***	0.168***	0.169***	0.168***
	(0.0320)	(0.0320)	(0.0266)	(0.0282)	(0.0290)
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	1112	1112	1112	1112	1112
Panel B: Privately Agree					
Treat	-0.00965	-0.0114	-0.0105	-0.0163	-0.0143
	(0.0416)	(0.0415)	(0.0403)	(0.0387)	(0.0379)
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	1260	1260	1260	1260	1260
Topic FE	✓	✓	✓	✓	✓
Baseline guesses		✓	✓	✓	✓
Session Controls			✓	✓	✓
Ind Controls				✓	✓

Standard errors clustered at the Zoom session level.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: the outcome variable $y_{i,j} = 1$ if individual i publicly express their views on topic j. Combined results on four topics (renaming schools affirmative action, death penalty and immunizations) are reported in this table. Session controls include time of the session, week of the session, group size and moderator FE. Individual controls include gender, race and ethnicity, year in school, major and self-reported political ideology.

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. Specifically, we estimate:

$$PublicDisagree_{i,j} = \beta_1 \%\tilde{Agree}^{M}_{i,j} + X_i + \theta_k + \varepsilon_{i,j} \tag{14}$$

where we use $Treat$ as an instrumental variable for $\%\tilde{Agree}^{M}_{i,j}$ and the first-stage is estimated by $\%\tilde{Agree}^{M}_{i,j} = \delta_1 Treat + X_i + \theta_k + \varepsilon_{i,j}$. Our 2SLS results indicate that on average, a 1 percentage point decrease in the perceived share of agreement in the midline survey corresponds to a 2.5% higher probability of publicly expressing a socially inappropriate view (see Table C.13 in Appendix C). Again as predicted, 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. Specifically, participants in the Treatment group infer that the socially appropriate view (Agree) is approximately 10% less prevalent compared to the Control group's beliefs at endline (see Columns (4) - (6) in Table 2), relative to a 6% difference between the Treatment and Control group at midline.[13] Taken together, our results are consistent with a compounding or self-reinforcing effect. 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 Zoom discussions, which further decreases the perceived popularity of the socially appropriate view at endline. As a result, participants in the Treatment group have more accurate beliefs about true public opinion at endline relative to participants in the Control group.

This dynamic belief-updating process is visualized in Figure 6, which illustrates the gradual divergence in perceived belief distribution 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 actually diverges further from the true belief distribution overtime, increasing their overestimation of the popularity of the socially appropriate view on average. For the topics on Affirmative Action and Renaming Schools, we even observe a flip around the 50/50 threshold at endline for the Treatment group relative to the Control group. At endline, Treatment participants accurately perceive that the previously believed socially inappropriate view is actually the majority view. For example, for the Renaming Schools topics, participants in the Control group continue to (inaccurately) believe that the majority of participants agree schools should be renamed, while those in the Treatment group correctly infer that a majority of participants actually believe schools should *not* be renamed.

[13]The treatment effect on endline beliefs is stronger than on midline beliefs, although not statistically significant. When comparing the coefficients of *Treat* in Columns (3) and (6) in Table 2, the p-value of the difference in coefficients is 0.14.

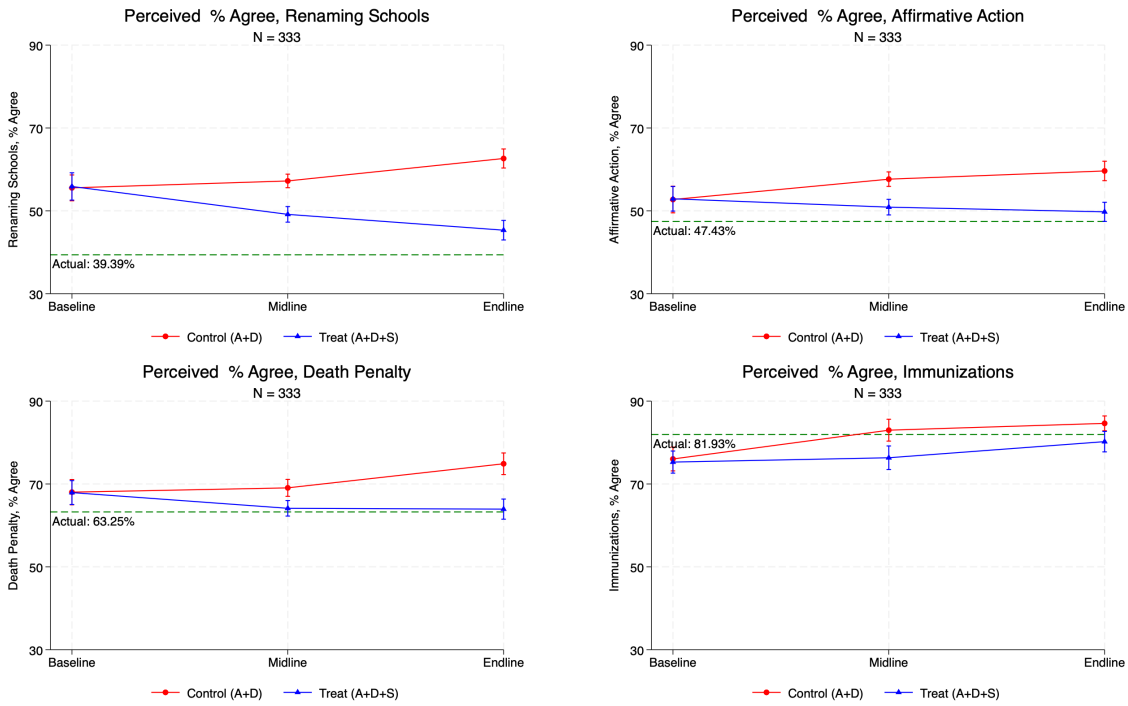


Figure 6: Perceived %*Agree* Over Time

Note: The graphs above show participants' average guesses about the share of agreement for each statement. Rounds 1, 2 and 3 correspond to beliefs elicited at baseline, midline, and endline respectively. The dashed green lines show the actual fraction of participants who privately agree with each statement. 95% confidence intervals are included.

For the Immunizations topic, recall that participants actually underestimated the popularity of the socially acceptable view at baseline. However, we observe similar patterns in inference and expression, as what the model predicts. First, individuals who hold the socially acceptable view are more likely to publicly express their views, although 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. Reassuringly, we observe no such dynamic effects for our placebo topic on Daylight Saving Time, consistent with social norms distorting expression and inference for socially sensitive topics but not for socially neutral topics.

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. In contrast, drawing attention to silence appears to effectively break 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 hold accurate beliefs about true public opinion while Control participants are trapped in echo chambers that further diverge from the truth.

5 Structural Estimates

Our experiment includes two rounds of public expression decisions and three rounds of belief elicitation. Yet an important insight that emerges from the reduced-form results in Section 4 is that the effects on inference and expression amplify and build on each other, which calls for an equilibrium analysis. To characterize equilibrium with different levels of attention to silence, we structurally estimate our model.

5.1 Structural Model Specification

Our structural model focuses on two primary variables: the level of attention to silence for participants in the Control and the Treatment groups, represented by λ_C and λ_T, respectively. By encompassing the dynamic nature of inference and expression, the model examines how these attention parameters affect the dual processes of belief-updating and public expression.

Concretely, we generate a simulated population with binary private beliefs $\theta_i \in \{A, D\}$ and an intrinsic value of speaking up that follows a normal distribution $V_i \sim N(v, \sigma_v^2)$. Individuals decide whether to publicly express their views based on their drawn V_i and the potential social sanction costs, determined by the topic-specific strengths of social sanction χ and the perceived prevalence of the socially appropriate view $E[\pi_i]_{t=0}$. In our benchmark model, we assume that the prior $\pi_{i,t=0} \overset{\text{iid}}{\sim} Beta(a, d)$. In the full model, we relax this assumption to accommodate heterogeneous priors: $\pi_{i,t=0} \sim (a_i, d_i)$, where $a_i \sim N(a, \sigma^2)$, $d_i \sim N(d, \sigma^2)$. Individuals with $\theta_i = A$ publicly express their opinions when their drawn V_i is greater than 0 while those with $\theta_i = D$ publicly express their opinions when their drawn V_i is greater than $\chi E[\pi_i]_{t=0}$.

Upon observing the expressed signals of first-movers $S_0 = \{S_{A,0}, S_{D,0}, S_{S,0}\}$, individuals update their beliefs about the prevalence of views following Equation 5. In the full model with heterogeneous priors, we assume that individuals project their own priors onto others when making inferences from their expressions and silence, $E[\pi_i(\lambda_{i,k})|S_0]_{t=1}$,

$k \in \{C, T\}$ for individuals in the Control Treatment group respectively. Note that in the full model, we also allow for heterogeneity in λ_C and λ_T, and $\lambda_{i,k}$ are drawn from $\lambda_{i,k} \sim N(\lambda_k, \sigma_\lambda^2)$ for $k \in \{C, T\}$. In the benchmark model, $a_i, d_i, E[\pi_i]_{t=0}, \lambda_{i,C}, \lambda_{i,T}$ are set to be constant across i.

With updated beliefs $\pi_{i,t=1}$, individuals make corresponding expression decisions in a second round following the same decision rule. After observing the expressed signals in the second round $S_k = \{S_{A,k}, S_{D,k}, S_{S,k}\}$, $k \in \{C, T\}$, individuals in the Control and Treatment group further update their beliefs $E[\pi_i(\lambda_{i,k})|S_k]_{t=2}$.

In our experiment, participants perceive signals in two different ways. The first set of signals $S_0 = \{S_{A,0}, S_{D,0}, S_{S,0}\}$ are presented through summary statistics in the midline survey, while the second set $S_k = \{S_{A,k}, S_{D,k}, S_{S,k}\}$ are conveyed in the Zoom sessions. Given the different formats, it is reasonable to assume that the levels of attention in these two rounds may vary. Therefore, we consider two specifications. In the benchmark model, λs are held constant throughout both belief-updating stages. In the extended model and the full model, λs are estimated separately for each stage.

To summarize, as outlined in the motivating framework and above, there are seven parameters that govern the expression decisions and the belief updating process for each topic. The parameters v and σ_v characterize the intrinsic value of honestly voicing one's opinions. The social sanction costs of expressing the socially inappropriate view are symbolized by χ. Parameters a and d determine the prior beliefs regarding support for A and D respectively. Different attention levels to silence in the Control and Treatment groups are represented by λ_C and λ_T respectively. Assuming that willingness to express and attention to silence are constant across the four topics, we have 16 parameters in total for the benchmark model. The extended model incorporates two more parameters to account for different λs across two stages. The full model introduces an additional three parameters $(\sigma, \sigma_{\lambda,1}, \sigma_{\lambda,2})$ to capture dispersion in priors and attention. We estimate these parameters with the following empirical moments, where each topic contributes nine empirical moments, adding up to 36 in total.

1. **Priors**: From the baseline survey, we elicit the perceived prevalence of agreement: $E[E[\pi_i]_{t=0}] = \bar{\pi}$.

2. **First Movers' Public Expressions**: In Zoom sessions with First Movers, we observe public expressions of socially appropriate and inappropriate views. Specifically, we observe the fraction $S_{A,0}/P_{A,0}$ and $S_{D,0}/P_{D,0}$, with $S_{A,0}$ and $S_{D,0}$ as the number of First Movers who publicly agreed or disagreed with each topic, and $P_{A,0}$, $P_{D,0}$ as the number who privately agreed or disagreed with each topic. These

fractions are matched with the simulated sample, following the expression rule: $express = 1$ if $V_i > 0$ for $\theta_i = A$ and $express = 1$ if $V_i > \chi E[\pi_i]_{t=0}$ for $\theta_i = D$.

3. **Midline Beliefs**: After observing a summary of the public expression of the First Movers ($S_{A,0}$, $S_{D,0}$ and $S_{S,0}$), Second Movers report their perceived belief distribution. $\bar{\pi}(\lambda_{i,k})_{t=1}$, $k \in \{C, T\}$ are the average perceived $\%Agree$ for the Control group and the Treatment group elicited in the midline survey.

4. **Second Movers' Public Expressions**: Two more moment conditions arise from the likelihood of expressing socially inappropriate views in Control and Treatment Zoom sessions: $S_{D,k}/P_{D,k}$, $k \in \{C, T\}$, where $S_{D,k}$ is the number of participants who publicly disagreed and $P_{D,k}$ is the number who privately disagreed for participants in group k (Control or Treatment). [14]

5. **Endline Beliefs**: The last two empirical moments are the perceived belief distributions elicited in the endline survey, matching $E[E[\pi_i(\lambda_{i,k})|S_k]_{t=2}] = \bar{\pi}(\lambda_{i,k})_{t=2}$.

To estimate the model, we employ a simulated method of moments estimator using the Markov Chain Monte Carlo (MCMC) algorithm. The vector of simulated moments as a function of parameters ξ is denoted by $m_N(\xi)$, while $\hat{m}$ represents the observed moments. We select parameters $\hat{\xi}$ that minimize the distance $(m_N(\xi) - \hat{m})' W (m_N(\xi) - \hat{m})$. We use the diagonal of the inverse of the variance-covariance matrix as the weighting matrix so that the estimator minimizes the sum of squared distances, weighted by the inverse variance of each moment. We include 20,000 individuals in the simulation. Table D.16 in Appendix D summarizes the simulated and empirical moments. Details about the simulation can be found in Appendix D.

5.2 Structural Estimation Results

In table 4, we report parameter estimates for three models. In the benchmark model, we assume homogeneous priors $\pi_i \overset{\text{iid}}{\sim} Beta(a, d)$ and homogeneous attention to silence within treatment assignment. In the extended model, we allow the attention parameters to vary across two stages. We use λ_1 to denote the level of attention that participants paid to the summary information provided about the public expression of First Movers, and λ_2 to denote the level of attention that participants paid to other Second Movers in

[14] We did not include the likelihood of expressing socially appropriate views as moment conditions because it does not rely on the perceived fraction of agreement. Therefore, these moments are co-linear with the moment conditions in 2.

their own Zoom sessions. In the full model, we further relax assumptions and allow for both heterogeneous priors and attention parameters.

The key parameters that we are interested in are the levels of attention to silence for the Control group and the Treatment group, denoted by $\hat{\lambda_C}$ and $\hat{\lambda_T}$. When assuming constant levels of attention at both stages, the average attention level among the Control group is estimated to be 0.23, while for the Treatment group it is noticeably higher at around 0.61.[15] If we allow attention parameters to vary across the two rounds, the findings become more nuanced. In the first round, when participants are presented with summary statistics that either exclude or include the number of silent First Movers, the average attention level to silence is approximately 0.25 for the Control group and 0.64 for the Treatment group. This difference in attention persists in the second round when participants infer from discussions in their own Zoom sessions. On average, the attention paid to silent Zoom participants is estimated to be around 0.35 for the Control group and 0.66 for the Treatment group. These estimates suggest that our Treatment information significantly increases attention to silence. Furthermore, this effect is not transient. Even as participants shift to new discussions and interpret new signals, those in the Treatment group continue to pay more attention to silence compared to Control participants.

The expression parameters reveal that the willingness of Berkeley college students to publicly express socially appropriate views follows a normal distribution with a mean of approximately -0.13. This translates to an average probability of 45% for publicly expressing one's views when they are socially acceptable.

We can also make some across-topic comparisons. Intuitively, the Beta distribution parameters, a and d, can be interpreted as "pseudo-counts" of agreement and disagreement, with larger a and d indicating higher certainty. It appears that respondents were most certain about the public opinion around the Death Penalty and Immunizations topics, consistent with our baseline survey results on participants' certainty about their guesses about the views of others.[16]

[15] If we reparameterize the model and use $\Delta_\lambda \equiv \lambda_T - \lambda_C$ to denote the difference in attention levels to silence among Control and Treatment participants. The 13% percentile of the posterior of Δ_λ, estimated by MCMC, is greater than 0.

[16] 83% of participants say that they are certain about their answers when guessing the belief distributions about immunizations, 65% for death penalty, 48% for affirmative action, and 40% for renaming schools.

Table 4: SMM: Estimating key parameters of the model and equilibrium beliefs

	Benchmark				Extended				Full			
Panel A: Attention Parameters												
	Same λ in Midline & Endline				λ_1 (Midline)		λ_2 (Endline)		λ_1 (Midline)		λ_2 (Endline)	
$\hat{\lambda_C}$	0.23				0.25		0.35		0.31		0.30	
	[0.06, 0.56]				[0.06, 0.58]		[0.10, 0.70]		[0.09, 0.65]		[0.08, 0.64]	
$\hat{\lambda_T}$	0.61				0.64		0.66		0.70		0.71	
	[0.27, 0.88]				[0.26, 0.89]		[0.30, 0.90]		[0.37, 0.92]		[0.40, 0.92]	
$\hat{\sigma_\lambda}$									0.48		0.49	
									[0.13, 1.13]		[0.15, 1.06]	
Panel B: Expression Parameters												
$\hat{v}$	-0.16				-0.15				-0.13			
	[-0.88, 0.37]				[-0.85, 0.37]				[-0.87, 0.40]			
Panel C: Topic-Specific Parameters												
	RS	AA	DP	Im	RS	AA	DP	Im	RS	AA	DP	Im
$\hat{\chi}$	2.25	2.28	1.95	2.21	2.18	2.28	2.06	2.07	1.99	2.04	1.97	2.14
	[0.74, 3.96]	[0.82, 4.03]	[0.61, 3.87]	[0.69, 4.08]	[0.77, 3.83]	[0.79, 4.05]	[0.66, 3.86]	[0.64, 3.80]	[0.73, 3.71]	[0.65, 3.81]	[0.64, 3.80]	[0.66, 3.97]
$\hat{a}$	4.68	5.00	5.82	6.57	4.42	4.57	5.46	6.41	4.16	4.35	5.04	5.81
	[2.50, 7.51]	[2.71, 7.93]	[3.32, 8.61]	[3.80, 9.01]	[2.45, 7.25]	[2.50, 7.64]	[3.14, 8.43]	[3.60, 8.84]	[2.36, 6.87]	[2.43, 6.91]	[2.90, 8.04]	[3.33, 8.43]
$\hat{d}$	4.01	4.18	2.96	2.72	3.67	3.95	2.92	2.75	3.47	3.50	2.71	2.68
	[2.11, 6.83]	[2.23, 6.93]	[1.67, 5.28]	[1.52, 4.80]	[2.01, 6.26]	[2.10, 6.67]	[1.62, 5.01]	[1.52, 4.69]	[1.93, 5.80]	[2.01, 5.79]	[1.55, 4.85]	[1.55, 4.86]
$\hat{\sigma}$									0.42			
									([0.14, 0.89]			
Weighted SSE	2.60				2.84				3.64			
Panel D: Beliefs in Equilibrium												
$E[\pi(\lambda = 0)^*]$	0.86	0.90	0.95	0.98	0.85	0.92	0.93	0.98	0.81	0.93	0.97	0.98
$E[\pi(\lambda = \hat{\lambda_C})^*]$	0.57	0.68	0.84	0.93	0.55	0.66	0.84	0.92	0.55	0.66	0.84	0.93
$E[\pi(\lambda = \hat{\lambda_T})^*]$	0.45	0.53	0.70	0.88	0.46	0.53	0.71	0.87	0.44	0.51	0.70	0.86
$E[\pi(\lambda = 1)^*]$	0.39	0.48	0.63	0.81	0.39	0.48	0.63	0.81	0.39	0.48	0.63	0.81

Notes: This table reports the estimated $\hat{\lambda_C}, \hat{\lambda_T}, \hat{v}, \hat{\sigma_v}, \hat{a}, \hat{d}, \hat{\chi}$ using the MCMC algorithm. We report the median as of posterior distributions of parameters as the parameter estimates, and the [16,84] credible intervals. The parameters $\hat{v}, \hat{\sigma_v}, \hat{\lambda_C}, \hat{\lambda_T}$ are assumed to be constant across the four topics ("RS": Renaming Schools; "AA": Affirmative Action; "DP": Death Penalty; "Im": Immunizations). In the benchmark model, we assume that a, b, λ_C, λ_T are the same for all individuals. In the extended model, we assume λ_C and λ_T are different in two stages, while holding the assumptions about homogeneous priors and attention levels. In the full model, we assume that a_i and d_i follow normal distributions $N(a, \sigma^2)$ and $N(d, \sigma^2)$, and $\lambda_{k,j,i}$ follow normal distributions $N(\lambda_{k,j}, \sigma^2_{\lambda_j})$ where $k \in \{C, T\}$, $j \in \{1, 2\}$. In panel D, we plug these parameters into Equations 8 and 9 to calculate the beliefs in equilibrium, assuming $\lambda = 0$ and $\lambda = 1$ respectively.

5.3 Beliefs in Equilibrium

With the estimated $\hat{v}, \hat{\sigma_v}, \hat{a}, \hat{d}, \hat{\chi}, \hat{\lambda_c}, \hat{\lambda_t}$, we can calculate beliefs in equilibrium, as formalized in Equations 8 and 9. Results are reported in Panel D in Table 4. For example, on Renaming Schools, if all individuals in the society pay full attention to silence, the perceived share of agreement will converge to the actual belief distribution: $E[\pi^*] = 0.39$. In contrast, if everyone is completely inattentive to silence, the perceived share of agreement will be above 0.86 in equilibrium.

The results on equilibrium beliefs suggest that the treatment effects we observe in our experiment are not transitory. Misperceived social norms could persist in the long run if individuals remain inattentive to silence. Because of these self-reinforcing effects, interventions that increases attention to silence could have long-lasting effects on beliefs. The treatment effects on beliefs are reinforced by expression and subsequent inference, resulting in divergent equilibrium outcomes.

6 Mechanisms

Section 4 establishes that providing participants with information that explicitly highlights the number of people who stay silent during a public discussion significantly changes subsequent inference and expression. In this section, we provide evidence consistent with *attention* to silence being a key mechanism driving these treatment effects.

6.1 Recall and Interpretation of Silence

In our endline survey, we measure participants' recall about the Zoom sessions they attended, including how many participants attended the 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 respectively.[17]

Our results indicate that Treatment individuals, who received information at midline intended to increase their attention to silence, have more accurate recall of the number

[17]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.

of participants who stayed silent during each Zoom discussion topic. On average, 63.7% of Treatment participants correctly recalled the number of participants who stayed silent during the Zoom discussion, compared to 49.3% of Control participants.[18] Furthermore, treatment effects are concentrated among students who correctly recall the number of participants who stayed silent during their Zoom discussions.

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 believe that silence is random, or uncorrelated with private beliefs, then a treatment that increases attention to silence should have no effect on subsequent inference or expression. If, instead, participants correctly infer the direction of selection bias into silence, namely that individuals who hold a socially inappropriate view are more likely to stay silent, 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. We classify a respondent as "sophisticated" if they correctly guess the direction of selection bias into silence. For example, for the Affirmative Action topic, 54% of respondents who stayed silent on the topic in the Zoom discussions held the socially inappropriate view. We categorize participants who estimated that more than half of the silent participants privately disagreed with the statement as sophisticated about the skewed selection into silence. Our results suggest that the treatment effects on endline beliefs are approximately 4% 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 C.14, Appendix C). [19]

6.2 Attention Versus Information Channels

Due to the nature of our experiment, Treatment participants receive both an information treatment intended to increase their attention to silence as well as new information on the public expression of their peers during their Zoom session. Recall that we con-

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

[19] 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% 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 (see Table C.14 in Appendix C). Table C.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.

ducted three rounds of belief elicitation for Second Movers: at baseline, midline (after participants received Control or Treatment information), and endline (after participating in Zoom discussions with other participants). The midline beliefs serve as immediate outcomes of our treatment and are easily interpreted as the direct effects of increased attention to silence. However, there are two factors that affect the belief-updating from midline to endline. First, as described in Section 4.3, the public expression that occurs in each Zoom session are different across Control and Treatment sessions. Second, participants in the Control and Treatment groups pay different levels of attention to silence in their own Zoom sessions (see Section 6.1 for suggestive evidence on accuracy of recall). In this section, we attempt to disentangle the effects of enhanced attention and new information in later rounds after the initial nudge treatment.

Specifically, we calculate two benchmarks for beliefs at endline using the structurally estimated parameters. The first benchmark holds information constant and asks what the Control group's endline beliefs would be if they had observed the signals in the Treatment Zoom sessions, given their attention level to silence. The second benchmark keeps the attention level constant and investigates what the Control group's endline beliefs would be if they paid the same level of attention to silence as the Treatment group, given the expressions they observed in the Control sessions. Results from our structural estimates suggest that both effects contribute roughly equally to our observed treatment effects, with around 40%-60% of the observed treatment effect explained by differences in attention to silence and around 50% of the treatment effect explained by differences in information between the Control and Treatment Zoom sessions. Results about these benchmarks can be found in Table C.15 in Appendix C.

7 Ecological Validity

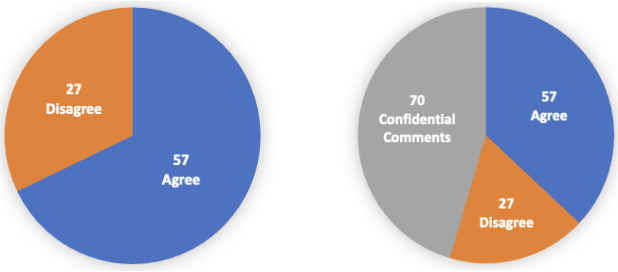
7.1 Berkeley Building Names

The renaming of school buildings is one of the topics that we discussed in our Zoom experiment, and it naturally extends to another experiment with Berkeley students. Using a similar experimental design, we explore how students update their beliefs based on actual signals generated on Berkeley's campus, and how that affects their actual political expression decisions, such as signing a petition.

The Berkeley Building Name Review Committee evaluates proposals submitted by university community members to potentially rename existing Berkeley buildings. Once a proposal is submitted, the committee gathers comments on the proposal. Community

members can share their comments publicly on the committee's website, or keep their comments confidential (for the committee only). For example, the committee collected 154 responses in total through their website's feedback form on one recent proposals to rename Moses Hall.

We conducted a brief survey with Berkeley students to understand their private views, perceptions about others' views and also public expression 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 randomly assign participants to either a Control group or a Treatment group where they see different pie charts summarizing the comments that the Building Name Review Committee collects, following the same design as the Zoom experiment. Specifically, the Treatment group's chart includes the number confidential comments, whereas the Control group's only shows the number of public comments saying "agree" and "disagree". Again both groups were informed of the total comment count, enabling them to calculate the number of confidential responses. Figure 7 presents the pie charts for the Control and Treatment groups respectively.



Control group Treatment group

Figure 7: Information Treatments, Berkeley Building Names

After presenting the pie charts, we incentivize students to guess the belief distribution by asking them *"Among the other Berkeley students who participate in this survey, what percentage do you think will privately answer "agree" and "disagree" with the following statement: 'All Berkeley buildings named after controversial historical figures should be renamed'"*. 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 named after controversial historical figures. The petition that participants get depends on their private beliefs. For example, those who privately agree 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".

The findings on inference and expression mirror those of our Zoom experiment. Specifically, participants in the Treatment group, who saw the pie chart including the confidential comments, estimated on average 8.5 percentage points fewer students would privately agree with renaming. Their guesses are thus closer to the actual share of agreement, at 48%. Among those who privately disagree, they are 20% more likely to sign a petition against renaming in the Treatment group relative to in the Control group. However, for those in agreement, the willingness to sign a petition in favor of renaming is not statistically different between the two groups. These results are detailed in Table 5.

The experiment with Berkeley Building Name Review Committee reveals that increasing attention to silence not only affects political discourse but also political action, such as the willingness to sign a petition.

Table 5: Guesses about %Agree, Willingness to Sign Petitions

	Guesses, %Agree		Pr(Sign|Disagree)		Pr(Sign|Agree)	
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	-8.963***	-8.429***	0.205***	0.208**	-0.0152	-0.0266
	(2.553)	(2.607)	(0.0733)	(0.0837)	(0.109)	(0.135)
Individual Controls		✓		✓		✓
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	169	169	169	169	169	169

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

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 petitions in favor of or against renaming. Individual Controls include gender, race and ethnicity, year in school, major, and self-reported political ideology.

7.2 Non-Responses and Misperceptions: ANES

In this section, we investigate the role of silence in political discourse and opinion climate with a broader spectrum of political issues, and with the general population.

Specifically, we test Proposition 3 about the correlation between silence and misperceptions with the American National Election Studies (ANES) data. 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 socially appropriate views could suppress the inappropriate views, and silence in turn can lead to less accurate beliefs about others.

The ANES, which is a representative national survey on public opinion and political involvement, has been conducted since 1948. It is conducted every two years before

2004, and every four years after that. 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.

Since 1970s, ANES elicit 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 questions: *Left to right; Liberal to conservative; Government services-spending; Defense spending; Government-private medical insurance, Guaranteed job income, Aid to Blacks.* Take the liberal-conservative scale as an example, the survey asks "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", and:

- Where would you place yourself on this scale?
- Where would you place the Democratic Party on this scale?
- Where would you place the Republican Party on this scale?

The detailed survey questions can be found in Appendix F.1. These questions allow us to construct misperception measures by comparing the actual ratings by Democratic survey respondents with the perceived ratings, similarly for the 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. Essentially, while the raw measure shows how off the perceptions are, the adjusted one also reveals if the deviation aligns with typical party norms. The adjusted measure is our preferred measure because it considers the direction of misperceptions.

To measure non-response rates, we code the dummy variable *non-response* as 1 if participants choose the following options when placing themselves on the 7-point scale: *Refused. Don't know. 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 rates as a lower bound of silence in public expression.

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 F.20, after controlling for year, topic and party fixed effets, as well as the

actual beliefs, a 10% increase in the non-response rate is correlated with approximately 2.6 percentage point increase in misperceptions. Table F.21 in the Appendix shows that this positive correlation also holds for the lag variable of *%No Response*, although the coefficients are generally smaller and not statistically significant.

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 the social norms, and conversely, when more people stay silent, people could have less accurate beliefs about others. In fact, our Zoom experiment indicates the plausibility of both scenarios. Instead of discussing the causal relationship, we highlight this positive correlation as predicted by Proposition 3 in the conceptual framework. This is not a trivial correlation: it would be absent if we didn't account for silence in political discourse, if silence wasn't 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.

Table 6: % No Response and Misperceptions, ANES

	Misperception (Raw)	Misperception (Raw)	Misperception (Raw)	Misperception (Adjusted)	Misperception (Adjusted)	Misperception (Adjusted)
Panel A: All Sample						
% No Response	0.0506	0.277***	0.162*	0.141**	0.405***	0.263**
	(0.0481)	(0.0744)	(0.0829)	(0.0633)	(0.108)	(0.117)
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
Year & Topic FE		✓	✓		✓	✓
Actual Beliefs			✓			✓
Party FE			✓			✓
Panel B: In-Person Surveys						
% No Response	0.0220	0.221***	0.127	0.105*	0.396***	0.202*
	(0.0470)	(0.0757)	(0.0846)	(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
Year & Topic FE		✓	✓		✓	✓
Actual Beliefs			✓			✓
Party FE			✓			✓

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: *Non-response* is a dummy variable that is equal to 1 if participants choose the following options when placing themselves on the 7-point scale: *Refused. Don't know. Haven't thought much about this.* Misperception (Raw) takes the absolute differences between the actual ratings by Democrats and perceived ratings by all respondents. Misperception (Adjusted) standardizes the direction so that positive misperceptions are consistent with stereotypes.

8 Conclusion

Disagreement is a healthy and vital part of public discourse and social progress. Yet disagreement can also feel challenging and uncomfortable. College campuses are ideally spaces that foster healthy debate and disagreement, particularly in service to learning

and innovation. Yet college also tends to be a time when individuals are particularly concerned about their social image.

In this paper, we study how social norms interact with public discourse on college campuses. We conduct experiments with UC Berkeley undergraduate students, a student body that is famously liberal. Our experiments reveal that social norms shape public discourse on campus. Students who hold views they perceive to be socially appropriate are more likely to express them to other students. Likewise, students who hold perceived socially inappropriate views are more likely to stay silent. After observing the public expression of other students, which is skewed towards the socially appropriate view, students further increase their beliefs about the popularity of the socially appropriate view, are 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 shows that inattention to silence is one 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 dissent 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 rich 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 spirals of silence, or more generally, phenomena where social norms diverge from majority private preferences. In doing so, our study also 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 or reads next to displays about the number of likes/dislikes and comments for each post. Likewise, when summarizing results from public discourse forums such as town hall meetings or public 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 social and political actions, the accuracy of those beliefs has important implications for civic engagement. 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

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.

Bifulco, Patrizio, Jochen Gluck, Oliver Krebs, and Bohdan Kukharskyy. 2022. "Single and Attractive: Uniqueness and Stability of Economic Equilibria under Monotonicity Assumptions." Papers 2209.02635, arXiv.org.

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

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.

Braghieri, Luca. 2021. "Political correctness, social image, and information transmission." *Work. Pap., Stanford Univ., Stanford, CA* .

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.

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, 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.

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`.

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.

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).

Gibbs, Jack P. 1965. "Norms: The Problem of Definition and Classification." *American Journal of Sociology* 70 (5):586–594.

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.

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.

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.

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.

Lapinski, Maria Knight and Rajiv N. Rimal. 2006. "An Explication of Social Norms." *Communication Theory* 15 (2):127–147.

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

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.

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/`.

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.

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.

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 for Proposition 2

(i) Full attention to Silence ($\lambda = 1$): p is the actual fraction of people with $\theta_i = A$. Assuming uniform mass and simplifying the notation: $E[\pi_1^*] \equiv E[\pi^*(\lambda = 1)]$, then as discussed in Section 2.1,

$$\begin{aligned} S_A^* &= [1 - G(0)]p \\ S_D^* &= [1 - G(\chi E[\pi_1^*])](1 - p) \\ S_S^* &= G(0)p + G(\chi E[\pi_1^*])(1 - p) \end{aligned}$$

Plugging in Equation 8:

$$\begin{aligned} E[\pi_1^*] = \frac{a}{a+d} &= \frac{S_A^* + S_S^* \cdot \frac{a}{a+d \cdot G(\chi E[\pi_1^*])/G(0)}}{S_A^* + S_D^* + S_S^*} \\ &= [1 - G(0)]p + [G(0)p + G(\chi E[\pi_1^*])(1 - p)] \frac{1}{1 + (1/E[\pi_1^*] - 1) \cdot G(\chi E[\pi_1^*])/G(0)} \\ &= [1 - G(0)]p + [G(0)p + G(\chi E[\pi_1^*])(1 - p)] \frac{G(0)E[\pi_1^*]}{G(0)E[\pi_1^*] + G(\chi E[\pi_1^*])(1 - E[\pi_1^*])} \end{aligned}$$

To further simplify notation, denote $y \equiv E[\pi_1^*]$, the equation above can be re-written as:

$$\begin{aligned} y &= [1 - G(0)]p + [G(0)p + G(\chi y)(1 - p)] \frac{G(0)y}{G(0)y + G(\chi y)(1 - y)} \\ \Leftrightarrow &(y - p)[G(0) + G(\chi y) - yG(\chi y) - G(0)G(\chi y)] = 0 \\ \Leftrightarrow &y = p \text{ or } y = 1 + G(0)(\frac{1}{G(\chi y)} - 1) > 1 \end{aligned}$$

$E[\pi_1^*] = y = p$ is the only solution for $y \in [0, 1]$. With full attention to silence, the perceived belief distribution is equl to the actual belief distribution in equilibrium.

(ii) Inattention to Silence ($\lambda = 0$):

$$\begin{aligned} E[\pi_0^*] = \frac{[1 - G(0)]p}{[1 - G(0)]p + [1 - G(\chi E[\pi_0^*])](1 - p)} &> p \\ \Leftrightarrow \frac{[1 - G(0)]}{[1 - G(0)]p + [1 - G(\chi E[\pi_0^*])](1 - p)} &> 1 \\ \Leftrightarrow G(\chi E[\pi_0^*]) &> G(0) \end{aligned}$$

This obviously holds with social saction costs $\chi > 0$.

(iii) Partial attention to silence ($0 < \lambda < 1$):

$$\begin{aligned} E[\pi^*] &= \frac{S_A^* + \lambda S_S^* \cdot \frac{a}{a+d \cdot G(\chi E[\pi^*])/G(0)}}{S_A^* + S_D^* + \lambda S_S^*} \\ &= \frac{[1-G(0)]p + \lambda[G(0)p + G(\chi E[\pi^*])(1-p)]\frac{G(0)E[\pi^*]}{G(0)E[\pi^*]+G(\chi E[\pi^*])(1-E[\pi^*])}}{1+(\lambda-1)[G(0)p + G(\chi E[\pi^*])(1-p)]} \end{aligned}$$

Next we prove that $E[\pi^*(\lambda)]$ is monotonic. Suppose that there exists $\lambda_1 \neq \lambda_2$ such that $E[\pi^*] = E[\pi^*(\lambda_1)] = E[\pi^*(\lambda_2)]$, we get

$$E[\pi^*] = \frac{[1-G(0)]p + \lambda_1[G(0)p + G(\chi E[\pi)^*])(1-p)]\frac{G(0)E[\pi^*]}{G(0)E[\pi^*]+G(\chi E[\pi^*])(1-E[\pi^*])}}{1+(\lambda_1-1)[G(0)p + G(\chi E[\pi^*])(1-p)]} \tag{15}$$

$$E[\pi^*] = \frac{[1-G(0)]p + \lambda_2[G(0)p + G(\chi E[\pi^*])(1-p)]\frac{G(0)E[\pi^*]}{G(0)E[\pi^*]+G(\chi E[\pi^*])(1-E[\pi^*])}}{1+(\lambda_2-1)[G(0)p + G(\chi E[\pi^*])(1-p)]} \tag{16}$$

Multiplying both sides with the denominator and then taking the difference of the two equations above, we can get

$$\Delta\lambda \cdot E[\pi^*] = \Delta\lambda \cdot \frac{G(0)E[\pi^*]}{G(0)E[\pi^*] + G(\chi E[\pi^*])(1-E[\pi^*])} \tag{17}$$

With the assumption that $\Delta\lambda \equiv \lambda_2 - \lambda_1 \neq 0$, Equation 17 always holds if and only if $\frac{G(0)}{G(0)E[\pi^*]+G(\chi E[\pi^*])(1-E[\pi^*])} = 1 \Leftrightarrow G(0) = G(\chi E[\pi^*])$, which contradicts with $G(0) < G(\chi E[\pi^*])$, $\chi > 0$.

Therefore, there does not exist $\lambda_1 \neq \lambda_2$ such that $E[\pi^*] = E[\pi^*(\lambda_1)] = E[\pi^*(\lambda_2)]$. $E[\pi^*(\lambda)]$ is monotonic in λ. We've already shown in (i) and (ii) that $E[\pi(0)^*] > E[\pi(1)^*] = p$, hence $E[\pi^*(\lambda)]$ decreases in λ, $\lambda \in [0,1]$.

A.2 Proof for Proposition 3

Positive Correlation between Silence and Misperceptions:
Silence and Misperceptions in period t:

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

Silence in period t and misperceptions in period t+1 (assuming uninformative prior):

$$
\begin{aligned}
Cov(S_S, \Delta|\mathcal{S}) &\propto Cov(\frac{[1-G(0)]p + \frac{\lambda S_S}{G(\chi E[\pi])}}{1+(\lambda-1)S_S}, S_S) \\
&= [1-G(0)]p \cdot Cov(\frac{1}{1+(\lambda-1)S_S}, S_S) + \lambda \cdot Cov(\frac{S_S}{G(\chi E[\pi])(1+(\lambda-1)S_S)}, S_S) \\
&> 0 \; for \; \lambda \in [0,1)
\end{aligned}
$$

A.3 Heterogeneous Priors

In this section we discuss heterogeneous priors about the belief distribution. Let's consider an individual i, whose priors about p follows a Beta distribution, denote by $\pi_i \sim Beta(a_i, d_i)$. Here, each individual is characterized by (V_i, π_i), where V_i denotes the individual's willingness to express a belief, and π_i describes the prior distribution. The expression rule for individual i is then determined by:

$$
\begin{aligned}
Pr(e_i = A|\theta_i = A) &= 1 - F(0) \\
Pr(e_i = D|\theta_i = D) &= 1 - F(\chi)
\end{aligned}
$$

Where F is the cdf of $V_i/E[\pi_i]$.

Individual i, with prior $\pi_i \sim Beta(a_i, d_i)$, observes signals S_A, S_D and S_S. Higher order beliefs about $E[\pi_j]$ will affect how individual i interpret signals. For simplicity, we assume that individual i projects his own prior $E[\pi_i]$ to all other individuals and updates his beliefs accordingly. His posteriors γ_i can be formulated as:

$$
E[\gamma_i(\lambda_i)] = \frac{a_i + S_A + \lambda_i S_S \cdot \frac{a_i}{a_i + d_i \cdot G(\chi E[\pi_i])/G(0)}}{a_i + d_i + S_A + S_D + \lambda_i S_S} \tag{18}
$$

Then, Proposition 1 holds with some slight modification:

Proposition 4. *If the socially appropriate views are expressed more often ($S_A \geq S_D$), for individuals with the same priors $\pi_i = \pi_j$, $E[\gamma_i(\lambda_i)] < E[\gamma_i(\lambda_i)]$ iff $\lambda_i > \lambda_j$.*

Intuitively, conditional on holding same priors, those who pay more attention to silence (reflected by a larger value of λ_i) will have lower estimates about the popularity of socially appropriate views.

We now turn to an examination of beliefs and expressions in equilibrium.

With full attention to silence: $\lambda_i = 1$, the equilibrium beliefs π_i^* satisfies:

$$\begin{aligned} E[\pi_i^*] = \frac{a_i}{a_i + d_i} &= \frac{S_A^* + S_S^* \cdot \frac{a_i}{a_i + d_i \cdot G(\chi E[\pi_i^*])/G(0)}}{S_A^* + S_D^* + S_S^*} \\ &= [1 - F(0)]p + \frac{F(0)p + F(\chi)(1-p)}{1 + (1/E[\pi_i^*] - 1) \cdot G(\chi E[\pi_i^*])/G(0)} \end{aligned} \tag{19}$$

Here, $p = E[\pi^*] = E[\pi_i^*]$ provides one solution to Equation 19. Our goal is to prove the uniqueness within the interval $(0, 1)$ of this solution. To simplify notation, we denote $y = E[\pi_i^*]$. Suppose that there exists multiple solutions y_i, then

$$\begin{aligned} p(Y) &= \sum_i \alpha_i \delta(Y - y_i) \\ F(\chi) &= \sum_i \alpha_i G(\chi y_i) \end{aligned}$$

Equation 19 can be re-written as:

$$M \equiv (y_i - p)K - G(0)L = 0$$

Where

$$\begin{aligned} K &= G(0)y_i + G(\chi y_i) - G(\chi y_i)y_i \\ L &= (1-p)\sum_j \alpha_j G(\chi y_j)y_i - pG(\chi y_i)(1 - y_i) \end{aligned}$$

To further simplify the analysis, we modify the assumption about G (cdf of V_i) and assume that it follows a uniform distribution: $G \sim U(-\chi, \chi)$.

$$M_i(y_1, ..., y_n) = 2(y_i - p)(1 + y_i - y_i^2) + p(1 - y_i^2) - (1-p)y_i \sum_j \alpha_j (1 + y_j)$$

For $y \in (0,1)$ and $p \in (0,1)$, $\frac{\partial M_i}{\partial y_i} = -3y_i^2 + (2+p)y_i + (1-p) + p\sum_j \alpha_j(1+y_j) > 0$, $\frac{\partial M_i}{\partial y_j} = py_i\alpha_j > 0$. The Jacobian matrix $|DM(y)|$ is irreducible for y. Therefore, with monotonicity in all dimensions in $(0,1)$, $y = p$ is the only solution in equilibrium. (Bifulco et al., 2022)

For individuals who are completely inattentive to silence ($\lambda = 0$), $y = E[\pi_i^*]$ satisfies:

$$y = \frac{[1 - F(0)]p}{[1 - F(0)]p + [1 - F(\chi)](1-p)} > p$$

The inequality above holds because $F(\chi) > F(0)$.

Finally, we use simulations to visually illustrate the divergent paths leading to equilibrium beliefs. Figure A.8 presents the evolution in beliefs for 1000 individuals over the course of 50 periods. In the scenario depicted, A is considered the socially appropriate view, but in reality, only 40% of people in favor of A. The red line represents the belief trajectory with complete inattention to silence ($\lambda = 0$), where we do observe the spiral of silence. In this case, the view A is perceived to become more and more popular over time, even though it is only supported by a minority. The blue line represents the belief updating path when individuals fully attend to silence, leading to gradual convergence to the actual belief distribution of 40% in favor of A. The green line demonstrates a case that lies in between, with partial attention to silence. Another observation from the graph is the narrowing of the standard deviation of beliefs within society over time. This trends indicates that, despite the heterogeneity in priors, the beliefs within the community tend to gradually converge towards the same equilibrium point, highlighting a collective alignment process in this dynamics.

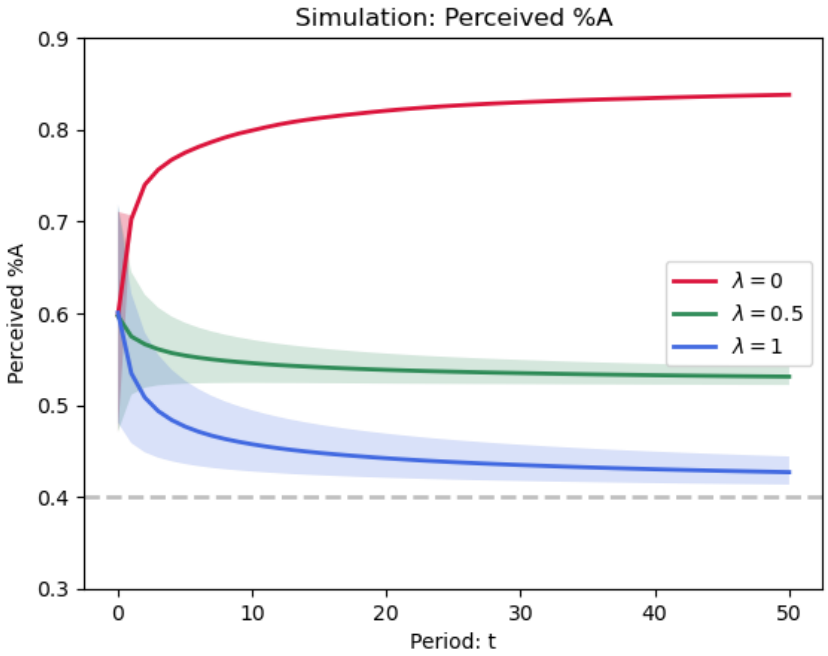


Figure A.8: Simulation, Perceived Fraction with $\theta = A$

Note: This figure shows the simulation with 1000 people over 50 periods. A is the socially appropriate view, and the actual fraction of people in favor of A is 40%. In each period, individuals make expression decisions, observe the expressed signals and update their beliefs about $\%A$. Individual i's prior about $\%A$ follows Beta distribution $Beta(a_i, d_i)$. To introduce heterogeneity in priors, we randomly draw a_i and d_i from $N(\bar{a}, \sigma^2)$ and $N(\bar{d}, \sigma^2)$, where $\bar{a} = 6, \bar{d} = 4, \sigma = 0.15$. V_i is drawn from the $N(0, 1)$ and the social sanction parameter χ is set to be 2.

Another type of heterogeneity in priors arises when we distinguish between individuals

who privately hold the socially appropriate or inappropriate views. In our experiment, we observe that people who privately hold the socially inappropriate views (D) have relatively lower estimates about the prevalence of socially appropriate views ($\%A$). To explore belief evolution over time, contingent on private beliefs and attention to silence, we run a simulation using the same parameters from Figure A.8. However, in this simulation we assume that people with $\theta = D$ begin with a more accurate prior about $\%A$. The simulation results are visualized in Figure A.9.

The red and dark blue lines illustrate the belief changes over time for those with $\theta = A$ when they pay no attention or complete attention to silence. In parallel, the orange and light blue lines show the belief evolution for those with $\theta = D$. We observe similar divergent patterns for individuals holding different private views, and thus having different priors. With no attention to silence, individuals with $\theta = D$ swiftly become influenced by the prevalent expressions favoring opinion A, despite initiating with a more accurate prior. After the initial several periods, their beliefs become identical with those with $\theta = A$ who start with higher estimates about $\%A$. In contrast, with full attention to silence, those who start with lower guesses about $\%A$ converge to the true value sooner. These trends not only align intuitively but also mirror the observations in our experiment.

A.4 Other Ways of Updating

We assumed Bayesian updating in the model. In this section, we examine other heuristic updating methods that participants might employ. For example, participants might project their prior beliefs onto silence, map the expressed signals on silence, or adopt a straightforward 50/50 rule. Figure A.10 displays the simulation results for these alternative updating rules. The red line shows the perceived $\%A$ progression over time when silence is completely ignored. The orange line depicts a scenario where people project the frequency of expressed opinions onto silence. As expressed signals lean towards the socially appropriate view A, mapping this to silence leads to heightened assumptions about the predominance of view A. The green line demonstrates the situation where people project their priors to silence. If there is an overestimation of the popularity of socially appropriate views, this bias persists and is not corrected over time. The purple line symbolizes the simple 50/50 rule, where people assume that half of the silent group hold the socially appropriate view, leading posterior beliefs to converge to a 50% estimation. Finally, the blue line shows the Bayesian updating results, which aligns most closely with the patterns we observed in the experiment.

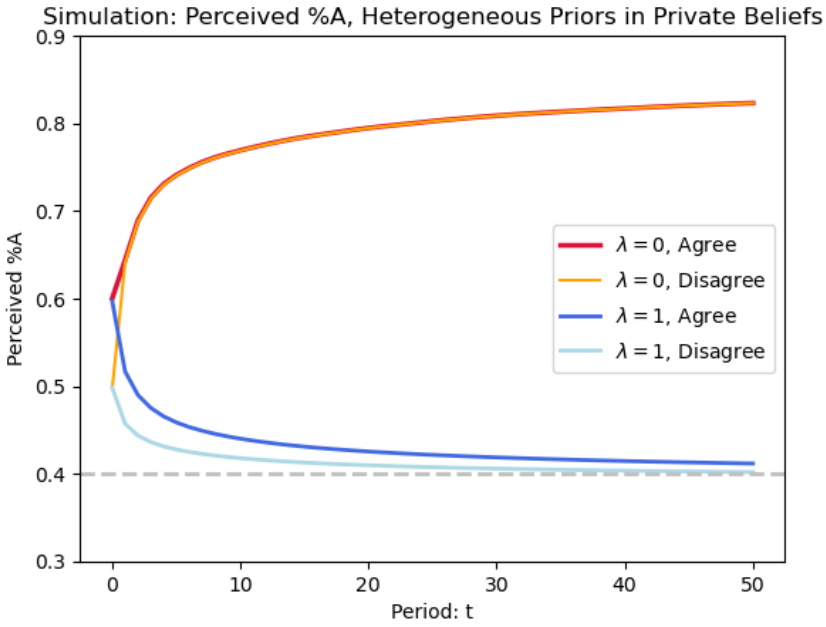


Figure A.9: Simulation, Perceived $\%Agree$, Heterogeneous Priors in Private Beliefs

Note: This figure shows the simulation with 1000 people over 50 periods. A is the socially appropriate view, and the actual fraction of people in favor of A is 40%. In each period, individuals make expression decisions, observe the expressed signals and update their beliefs about $\%A$. Individual i's prior about $\%A$ follows Beta distribution $Beta(a_i, d_i)$. To introduce heterogeneity in priors, we randomly draw a_i and d_i from $N(\bar{a}, \sigma^2)$ and $N(\bar{d}, \sigma^2)$, where $sigma = 0.15$, $\bar{a} = 6, \bar{d} = 4$ for those with $\theta = A$ and $\bar{a} = 5, \bar{d} = 5$ for those with $\theta = D$. V_i is drawn from the $N(0, 1)$ and the social sanction parameter χ is set to be 2.

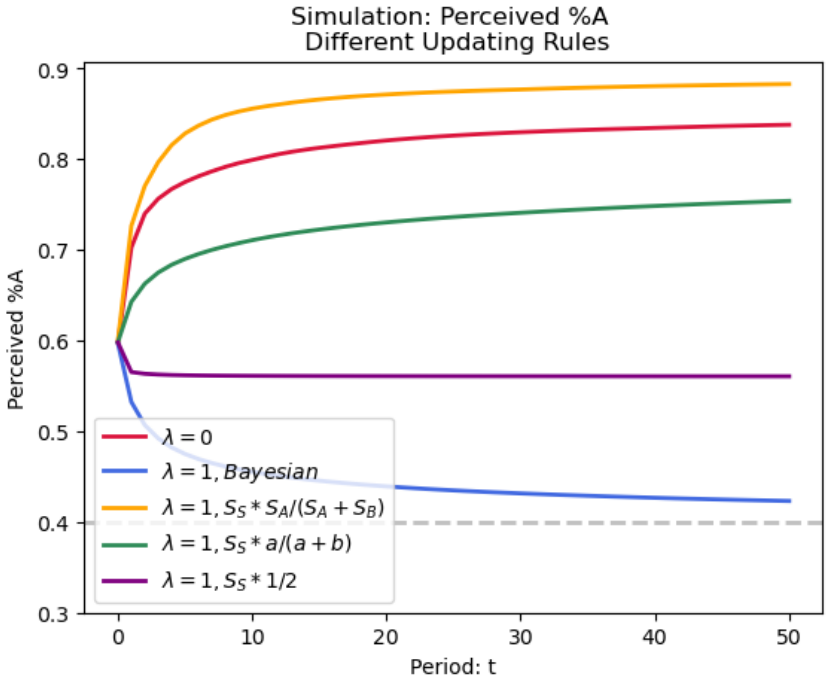


Figure A.10: Simulation, Perceived Fraction with $\theta = A$

Note: This figure shows the simulation with 1000 people over 50 periods, using different updating rules. A is the socially appropriate view, and the actual fraction of people in favor of A is 40%. In each period, individuals make expression decisions, observe the expressed signals and update their beliefs about $\%A$. Individual i's prior about $\%A$ follows Beta distribution $Beta(a_i, d_i)$. To introduce heterogeneity in priors, we randomly draw a_i and d_i from $N(\bar{a}, \sigma^2)$ and $N(\bar{d}, \sigma^2)$, where $\bar{a} = 6, \bar{d} = 4, \sigma = 0.15$. V_i is drawn from the $N(0, 1)$ and the social sanction parameter χ is set to be 2.

A.5 Extensions: Multiple Categories

The model can be generalized to accommodate multiple categories of private preferences. Specifically, assuming there are $K \geq 2$ categories of private preferences $\theta_i \in \{A_1, A_2, ..., A_K\}$ and the prior about the belief distribution follows Dirichlet distribution $\pi \sim Dir(K, \alpha)$, where $f(\pi_1, ...\pi_K; \alpha_1, ..., \alpha_K) = \frac{1}{B(\alpha)} \Pi_{i=1}^{K} \pi_i^{\alpha_i - 1}$.

Similarly, we assume that the utility of honestly voicing one's opinions $V_i \sim N(v, \sigma_v^2)$. Therefore, the probability of expressing for those with $\theta_i = A_k \in \mathbf{A_0}$ is $1 - G(\chi \sum_{j \in \mathbf{A_1}} E[\pi_j])$ if A_k is in the set of socially inappropriate views $\mathbf{A_0}$. The social sanction costs increase in the perceived popularity of the socially appropriate views $\mathbf{A_1}$.

After observing expressed signals $\mathbf{S} = (S_1, ..., S_K)$ and silence S_S, individuals form posteriors about the opinion climate:

$$\gamma_\lambda | S, \alpha \sim Dir(K, f(\alpha, s, \lambda))$$

where

$$f(\alpha, s, \lambda) = \begin{cases} \alpha_k + s_k + \lambda \frac{\alpha_k G(\chi \sum\limits_{j \in \mathbf{A_i}} E[\pi_j]) \cdot S_S}{\sum\limits_{k \in \mathbf{A_0}} \alpha_k G(\chi \sum\limits_{j \in \mathbf{A_i}} E[\pi_j]) + \sum\limits_{j \in \mathbf{A_1}} \alpha_j G(0)} \equiv \alpha_k + s_k + \lambda X_k & k \in \mathbf{A_0} \\ \alpha_j + s_j + \lambda \frac{\alpha_j G(0) \cdot S_S}{\sum\limits_{k \in \mathbf{A_0}} \alpha_k G(\chi \sum\limits_{j \in \mathbf{A_i}} E[\pi_j]) + \sum\limits_{j \in \mathbf{A_1}} \alpha_j G(0)} \equiv \alpha_j + s_j + \lambda X_j & j \in \mathbf{A_1} \end{cases} \tag{20}$$

$$E[\gamma_\lambda^k] = \frac{\alpha_k + S_k + \lambda X_k S_S}{\sum \alpha_k + \sum S_k + \lambda S_S} \tag{21}$$

Take FOC w.r.t the attention parameter λ, we can get $\partial E[\gamma_\lambda^k]/\partial \lambda > 0$, for $k \in \mathbf{A_0}$ and $\partial E[\gamma_\lambda^j]/\partial \lambda < 0$, for $j \in \mathbf{A_1}$ under the same assumptions of Proposition 1.

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 Discussion

Figure B.11 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

Table B.7: Socially Appropriateness of each Statement

Topic	Very Inappropriate	Somewhat Inappropriate	Neutral	Somewhat Appropriate	Very Appriproate
Affirmative Action	6.23	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 modal response to each topic is indicated in bold.

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.

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

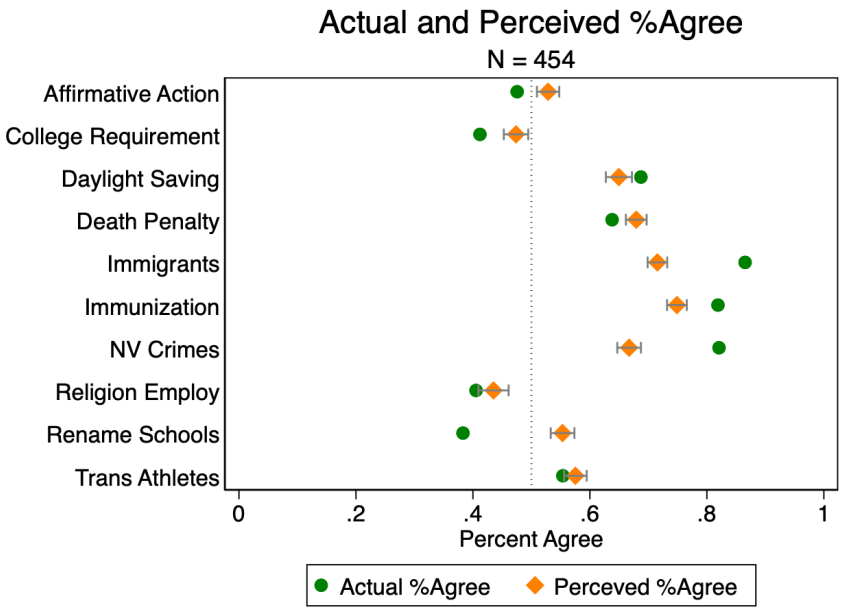


Figure B.11: Baseline Actual and Perceived Fraction of Student Agreement

did not appear to be socially sensitive. 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 Graphs and Tables

Table C.8: 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: 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.9: Attrition Table

	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	0.39	0.49	0.39	0.49	0.41	0.49	0.80
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	0.69	0.46	0.67	0.47	0.76	0.43	0.15
Panel C: Baseline Guesses							
Rename Schools	55.79	21.64	55.60	21.24	56.82	23.82	0.66
Affirmative Action	52.85	20.84	52.55	20.80	54.46	21.10	0.48
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	64.60	24.60	63.78	25.01	69.04	21.92	0.10
Panel D: Treatment Assignment							
treat	0.50	0.50	0.50	0.50	0.49	0.50	0.89

Notes: 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.10: Summary Statistics of First Movers

	Agree	Disagree	Silent	Total
Renaming Schools	7	5	13	25
Affirmative Action	7	5	13	25
Death Penalty	9	4	12	25
Immunizations	13	1	11	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.11: Summary Statistics of All First Movers

	Agree	Disagree	Silent	Total
Renaming Schools	10	10	30	50
Affirmative Action	11	9	30	50
Death Penalty	18	8	24	50
Immunizations	21	1	28	50

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 2 sessions to avoid overly-shifting Second Movers' beliefs in the midline survey.

Table C.12: Logit: $y_{i.j} = 1$ if individual i publicly express their views on topic j

	Affirmative Action	Death Penalty	Immunizations	Renaming Schools
Panel A: Privately Disagree				
Treat	0.0826**	0.258***	0.206	0.171***
	(0.0397)	(0.0563)	(0.129)	(0.0450)
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	-0.475	0.0916	0.187	-0.501
	(0.391)	(0.280)	(0.222)	(0.379)
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	✓	✓	✓	✓

Standard errors clustered at the Zoom session level.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: the outcome variable $y_{i,t} = 1$ if individual i publicly express their views on topic j. Session controls include time of the session, week of the session, group size and moderator FE. Individual controls include gender, year in school, race and ethnicity, major and self-reported political ideology.

Table C.13: 2SLS: $y_{i.j} = 1$ if individual i publicly express their views on topic j

	Express = 1	Express = 1	Express = 1	Express = 1
Panel A: Privately Disagree				
$\%\hat{Agree}^{M}_{i,j}$	-2.256***	-2.309***	-2.367***	-2.492***
	(0.549)	(0.529)	(0.487)	(0.530)
Mean	0.164	0.164	0.164	0.164
SD	0.371	0.371	0.371	0.371
IDs	278	278	278	278
Obs	1112	1112	1112	1112
First-Stage F stats		10.12		
Panel B: Privately Agree				
$\%\hat{Agree}^{M}_{i,j}$	0.160	0.184	0.170	0.287
	(0.679)	(0.663)	(0.634)	(0.658)
Mean	0.407	0.407	0.407	0.407
SD	0.492	0.492	0.492	0.492
IDs	315	315	315	315
Obs	1260	1260	1260	1260
First-Stage F stats		19.54		
Topic FE	✓	✓	✓	✓
Baseline guesses		✓	✓	✓
Session Controls			✓	✓
Ind Controls				✓

Standard errors clustered at the Zoom session level.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: the outcome variable $y_{i,t} = 1$ if individual i publicly express their views on topic j. Treat = 1 is the IV for $\%\hat{Agree}^{M}_{i,j}$. Combined results on four topics (renaming schools affirmative action, death penalty and immunizations) are reported in this table. Session controls include time of the session, week of the session, group size and moderator FE. Individual controls include gender, year in school, race and ethnicity, major and self-reported political ideology.

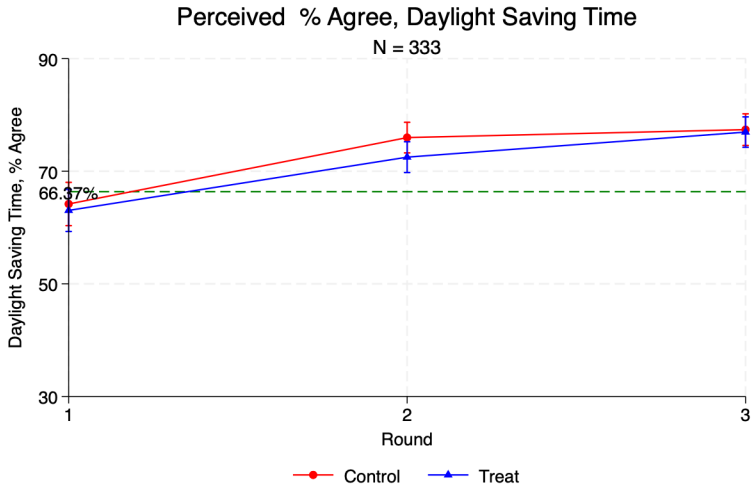


Figure C.12: Perceived $\%Agree$ Over Time, DST

Note: The graphs show participants' average guesses about the share of agreement. Round 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. 95% confidence intervals are included in the graph.

Table C.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***	-5.539***	-7.065***	-8.591***	-6.209***	-6.626***	-6.446***	-7.038***	-9.489***
	(1.357)	(1.177)	(1.752)	(1.418)	(1.070)	(1.862)	(1.282)	(1.161)	(3.268)
Treat x Private Agree	1.084								0.735
	(1.528)								(1.591)
Treat x Soph about Silence		-2.325							-2.235
		(1.567)							(1.596)
Treat x Female			0.506						0.768
			(2.040)						(2.018)
Treat x Asian				3.335*					3.924*
				(1.912)					(2.311)
Treat x White					-2.324				0.355
					(2.338)				(2.793)
Treat x Liberal						-0.105			0.471
						(2.146)			(2.198)
Treat x Senior							-0.559		-0.498
							(1.909)		(1.856)
Treat x Social Science								1.287	2.050
								(1.940)	(1.991)
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***	-7.965***	-11.27***	-11.82***	-9.938***	-10.70***	-10.88***	-10.00***	-13.41***
	(1.284)	(1.623)	(1.914)	(1.550)	(1.259)	(2.230)	(1.381)	(1.246)	(3.704)
Treat x Private Agree	6.099***								4.776***
	(1.706)								(1.464)
Treat x Soph about Silence		-4.193*							-3.317
		(2.434)							(2.459)
Treat x Female			1.225						1.031
			(2.097)						(1.953)
Treat x Asian				2.499					2.164
				(1.758)					(2.578)
Treat x White					-2.201				-0.235
					(2.067)				(2.991)
Treat x Liberal						0.409			0.124
						(2.787)			(2.257)
Treat x Senior							1.063		1.671
							(1.921)		(1.769)
Treat x Social Science								-1.530	-1.404
								(1.605)	(1.654)
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

Standard errors are clustered at the individual level.

Notes: the outcome variables are perceived %*Agree* elicited in the midline survey and he endline survey. Combined results on the four topics (renaming schools affirmative action, death penalty and immunizations) are reported. Standard errors 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 FE. Individual controls include gender, year in school, race and ethnicity, major and self-reported political ideology.

Table C.15: Endline Beliefs: Actual Beliefs and Benchmarks

	Renaming Schools	Affirmative Action	Death Penalty	Immunizations
Benchmark 1	0.52	0.53	0.67	0.82
Benchmark 2	0.54	0.54	0.68	0.83
Control, Endline	0.63	0.60	0.75	0.85
Treat, Endline	0.45	0.50	0.64	0.80

Notes: This table reports two benchmarks of the endline beliefs to disentangle the treatment effects. Benchmark 1 is calculated assuming that the Control group receives the publicly expressed signals in Treatment Zoom sessions. Benchmark 2 assumes that the Control group pays the same level of attention to Zoom discussions as the Treatment group.

Appendix D SMM Details

To estimate the model, we employ a simulated method of moments estimator using the Markov Chain Monte Carlo (MCMC) algorithm, implemented through the python package *emcee*. We include 20,000 individuals in the simulation and match the 36 empirical moments, setting 500 walkers in the parameter space and $nsteps = 2,000$, with a 10% burn-in period. In the benchmark model, we set the following bounds for the parameters $\xi = \{a_1, b_1, a_2, b_2, a_3, b_3, a_4, b_4, v, \chi_1, \chi_2, \chi_3, \chi_4, \lambda_C, \lambda_T\}$: $\{(1, 10), (1, 10), (1, 10), (1, 10), (1, 10), (1, 10), (1, 10), (1, 10), (-3, 3), (0.01, 5), (0.01, 5), (0.01, 5), (0.01, 5), (0, 1), (0, 1)\}$. [20] The full model adds bounds for $(\sigma, \sigma_{\lambda,1}, \sigma_{\lambda,2})$ as $(0.01, 2)$. The priors of parameters are assumed to be uniform within bounds. The initial guesses are (3, 2, 3, 2, 4, 2, 5, 2, 0.0, 2.0, 2.0, 2.0, 2.0, 0.2, 0.8) in the benchmark model, and (0.1, 0.2, 0.2) for $(\sigma, \sigma_{\lambda,1}, \sigma_{\lambda,2})$ in the full model. We experimented with different initial guesses and the final estimates are not the sensitive to different initial guesses. The estimates are also not sensitive to the number of walkers or the number of steps. We experimented with 400 walkers, 600 walkers, 2,000 steps and 3,000 steps and get similar results, suggesting that the parameter estimates already converge.

The MCMC run yields posterior distributions of parameters ξ. The corner plot below shows the posterior distributions for the benchmark model. In table 4, we report the median as the parameter estimates, along with the credible intervals of the posteriors.

[20] σ_v is set to be 1 to address the degeneracy in correlated moment conditions.

Table D.16: SMM: Simulated and Empirical Moments

	Simulated Moments			Empirical Moments	
	Parameter Assumptions	Moments	=	Moments	Calculated From
Priors (Baseline Survey)	$\pi_i \sim Beta(a_i, d_i)$	$E[E[\pi_i]_{t=0}]$	=	$\bar{\pi}_{t=0}$	Mean Guesses in Baseline Survey
First Movers Expression (A)	$e_i = A$ if $V_i > 0$	$\sum \mathbf{1}_{\{e_i=A\}} / \sum \mathbf{1}_{\{\theta_i=A\}}$	=	$S_{A,0}/P_{A,0}$	FM Zoom Sessions
First Movers Expression (D)	$e_i = D$ if $V_i > \chi E[\pi_i]_{t=0}$	$\sum \mathbf{1}_{\{e_i=D\}} / \sum \mathbf{1}_{\{\theta_i=D\}}$	=	$S_{D,0}/P_{D,0}$	FM Zoom Sessions
Midline Guesses (C)	$\pi_{i,t=1}(\lambda_C)\|S_0$ based on Eq 5	$E[E[\pi_i(\lambda_C)\|S_0]_{t=1}]$	=	$\bar{\pi}_{C,t=1}$	Mean Guesses in Midline Survey, Control
Midline Guesses (T)	$\pi_{i,t=1}(\lambda_T)\|S_0$ based on Eq 5	$E[E[\pi_i(\lambda_T)\|S_0]_{t=1}]$	=	$\bar{\pi}_{T,t=1}$	Mean Guesses in Midline Survey, Treat
Second Movers Expression (C)	$e_i = D$ if $V_i > \chi E[\pi_i(\lambda_C)]_{t=1}$	$\sum \mathbf{1}_{\{e_i=D\}} / \sum \mathbf{1}_{\{\theta_i=D\}}$	=	$S_{D,C}/P_{D,C}$	Control Zoom Sessions
Second Movers Expression (T)	$e_i = D$ if $V_i > \chi E[\pi_i(\lambda_T)]_{t=1}$	$\sum \mathbf{1}_{\{e_i=D\}} / \sum \mathbf{1}_{\{\theta_i=D\}}$	=	$S_{D,T}/P_{D,T}$	Treatment Zoom Sessions
Endline Guesses (C)	$\pi_{i,t=2}(\lambda_C)\|S_C$ based on Eq 5	$E[E[\pi_i(\lambda_C)\|S_C]_{t=2}]$	=	$\bar{\pi}_{C,t=2}$	Mean Guesses in Endline Survey, Control
Endline Guesses (T)	$\pi_{i,t=2}(\lambda_T)\|S_T$ based on Eq 5	$E[E[\pi_i(\lambda_C)\|S_T]_{t=2}]$	=	$\bar{\pi}_{T,t=2}$	Mean Guesses in Endline Survey, Treat

This table summarizes the simulated and empirical moments, along with explanations of the relevant parameters. In general, π_t indicates perceived belief distribution in different rounds. $e_i = \{A, 0, D\}$ indicates the expression decisions. V_i is the intrinsic value of expression, χ is the parameter for strength of social sanction. λ_C, λ_T represent attention to silence in Control and Treatment groups respectively. $S_{A,t}$ and $S_{D,t}$ are the number of participants who publicly agreed or disagreed with each topic in Zoom sessions. $P_{A,t}$ and $P_{D,t}$ are the number of participants who privately agreed or disagree with each topic.

Appendix E Silence and Polarization

E.1 Experimental Design

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 Zoom 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 of the statements.

Following the same procedure as the Zoom experiment, we randomly assign participants who complete the baseline survey into the First Movers group, the Control group or the 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 choose from the 5-point scale and to provide a brief explanation for their answers. They are informed that their responses will be shared publicly with other participants. Specifically, they read in the survey 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"*

Similarly as the Zoom experiment, we experimentally vary the salience of silence when presenting summary statistics about First Movers' public expression to Second Movers in the follow-up survey. The Control group sees pie charts excluding First Movers who skipped the public expression questions. The Treatment group sees pie charts including First Movers who skipped the questions. Second Movers then guess the percentage of Berkeley participants who would privately choose each option, ranging from "strongly agree" to "strongly disagree". Second Movers are also asked to choose from the 5-point scale and to provide a brief explanation for their answers, which are made public to other participants.

E.2 Results

Expression by First Movers

Public expression is selected not only based on socially appropriateness of opinions, but also the strength of opinions. Table E.17 present 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 holding socially acceptable views "agree" (selecting 4 or 5 in the baseline survey) are more likely to publicly express their views in the survey. Moreover, Participants holding extreme views (selecting 1 or 5) are around 20% more likely to publicly express them. Taken together, the public expression is skewed in the direction of socially appropriate views, and are more heavy-tailed than the actual belief distribution.

Table E.17: Expression Decisions by First Movers

	(1)	(2)	(3)
	Express =1	Express =1	Express = 1
Private Agree	0.339***	0.304***	0.271**
	(0.0935)	(0.0943)	(0.100)
Private Extreme	0.211**	0.203**	0.201**
	(0.0797)	(0.0782)	(0.0863)
Topic FE	✓	✓	✓
Baseline guesses		✓	✓
Demographics			✓
Mean	0.614	0.614	0.614
SD	0.489	0.489	0.489
IDs	35	35	35

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports the public expression decisions by First Movers on four topics (Affirmative Action, Death Penalty, Immunizations, and Renaming Schools). $y_{i,j} = 1$ 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.

Inference by Second Movers

The results on inference replicate what we found in the Zoom experiment, and show that experimentally drawing attention to silence reduces perceived polarization. Table E.18 reports Second Movers' perceived share of agreement, share of disagreement, and share of extreme views after viewing pie charts excluding or including First Movers who skipped the public expression questions. We find that, treatment participants have lower estimate of share of agreement, which is consistent with the Zoom experiment. Moreover, treatment participants have lower guesses about the prevalence of extreme views, suggesting that increased attention to silence leads to lower perceived polarization.

Expression by Second Movers

Table E.18: Control/Treatment Guesses

	(1)	(2)	(3)	(4)	(5)	(6)
	%Agree	*%Agree*	*%Disagree*	*%Disagree*	*%Extreme*	*%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

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports Second Movers' inference about the belief distribution. Individual controls include gender, race and ethnicity, year in school, major, and self-reported political ideology.

Finally, we also elicit Second Movers' willingness to publicly express their opinions and find similar dynamic patterns as in the Zoom experiment. Table E.19 demonstrates the public expression decisions by Second Movers who privately disagree with the socially appropriate views (selecting 1 or 2 in the baseline survey), and who privately hold moderate views (selecting 2, 3, or 4). The positive coefficients of *Treat* suggest 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.19: Control/Treat Expressions

	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

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports the public expression decisions by Second Movers on four topics (Affirmative Action, Death Penalty, Immunizations, and Renaming Schools). $y_{i,j} = 1$ 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.

Appendix F ANES Data

F.1 Survey Questions

Descriptions:

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?

F.2 Regression Tables: %No Response and Misperceptions

Table F.20: % No Response and Misperceptions, ANES

	Misperception (Raw)	Misperception (Raw)	Misperception (Raw)	Misperception (Adjusted)	Misperception (Adjusted)	Misperception (Adjusted)
Panel A: All Sample						
% No Response	0.0506	0.277***	0.162*	0.141**	0.405***	0.263**
	(0.0481)	(0.0744)	(0.0829)	(0.0633)	(0.108)	(0.117)
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
Year & Topic FE		✓	✓		✓	✓
Actual Beliefs			✓			✓
Party FE			✓			✓
Panel B: In-Person Surveys						
% No Response	0.0220	0.221***	0.127	0.105*	0.396***	0.202*
	(0.0470)	(0.0757)	(0.0846)	(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
Year & Topic FE		✓	✓		✓	✓
Actual Beliefs			✓			✓
Party FE			✓			✓

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: *non-response* is a dummy variable that is equal to 1 if participants choose the following options when placing themselves on the 7-point scale: *Refused. Don't know. Haven't thought much about this.* Misperception (Raw) takes the absolute differences between the actual ratings by Democrats and perceived ratings by all respondents. Misperception (Adjusted) standardizes the direction so that positive misperceptions are consistent with stereotypes.

Table F.21: Lag %No Response and Misperceptions, ANES

	Misperception (Raw)	Misperception (Raw)	Misperception (Raw)	Misperception (Adjusted)	Misperception (Adjusted)	Misperception (Adjusted)
Panel A: All Sample						
Lag % No Response	0.00813	0.126	0.0413	0.0709	0.227**	0.0550
	(0.0482)	(0.0766)	(0.0833)	(0.0633)	(0.110)	(0.117)
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	170	170	170	170	170	170
Year & Topic FE		✓	✓		✓	✓
Actual Beliefs			✓			✓
Party FE			✓			✓
Panel B: In-Person Surveys						
Lag % No Response	0.0116	0.158*	0.0794	0.0656	0.285**	0.118
	(0.0489)	(0.0799)	(0.0904)	(0.0642)	(0.114)	(0.126)
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	170	170	170	170	170	170
Year & Topic FE		✓	✓		✓	✓
Actual Beliefs			✓			✓
Party FE			✓			✓

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: *non-response* is a dummy variable that is equal to 1 if participants choose the following options when placing themselves on the 7-point scale: *Refused. Don't know. Haven't thought much about this.* Misperception (Raw) takes the absolute differences between the actual ratings by Democrats and perceived ratings by all respondents. Misperception (Adjusted) standardizes the direction so that positive misperceptions are consistent with stereotypes.

Appendix G Survey Instructions

G.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)

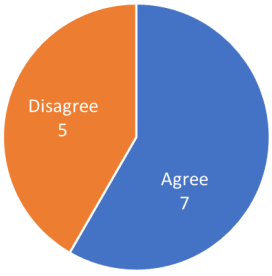
G.2 Midline Survey

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

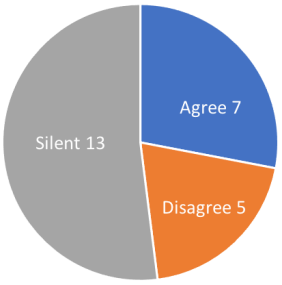
[**Welcome**] Thank you for completing the first part of our study. To help you get a better sense about the Zoom discussion, below we will show you summaries about earlier Zoom sessions conducted with other study participants who are also Berkeley students.

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)

G.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)

G.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.

G.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 H 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]