Threshold Disclosure in Collective Decisions*

Luca Braghieri[†] Leonardo Bursztyn[‡] Jan Fasnacht[§]

January 2026

Abstract

Voting-based collective decisions are typically made either anonymously or publicly. Anonymous voting protects truthful expression but conceals individual behavior; public voting provides information about individual votes, but, when one option is socially stigmatized, it can distort participation and choices. We introduce threshold majority voting, in which voters choose a disclosure threshold determining whether and when their votes are revealed. In an experiment at UC Berkeley on the participation of transgender women in women's sports, public voting nearly doubles abstention and reduces support for the stigmatized option. Threshold voting eliminates these distortions while revealing one-third of individual votes.

Keywords: preference falsification, voting mechanisms, social image, information transmission, institutional design

JEL Codes: D72, D82, C93

[*] The research described in this article was approved by the Committee for Protection of Human Subjects (CPHS) at the University of California, Berkeley (#00006252). This study was pre-registered in the AEA RCT Registry (AEARCTR-16968). We thank Don Moore, Aileen Wu, and Rowilma del Castillo for support at UC Berkeley. We are grateful to Matthew Gentzkow, Timm Gries, Massimo Morelli, and Guido Tabellini for helpful comments. We thank Heitor Sousa Borges for excellent research assistance.

[†] Bocconi University and NHH, CEPR, CESifo, and IGIER. (luca.braghieri@unibocconi.it)

[‡] University of Chicago and NBER. (lbursztyn@uchicago.edu)

[§] University of Chicago. (jfasnacht@uchicago.edu)

Collective decisions made through voting are typically conducted either anonymously or publicly. Anonymous voting encourages honest expression by shielding individuals from social scrutiny, but it obscures who supported which option. Public voting helps assign responsibility for the implemented outcome and facilitates downstream coordination by linking votes to identities, but, when one option is socially stigmatized, it can distort participation and expressed preferences.

Consider, as an example, a university committee voting on whether to expand diversity, equity, and inclusion (DEI) programs. In a public vote, social-image concerns may lead some members to support the proposal even if they privately disagree with it, thus distorting the outcome. In an anonymous vote, in contrast, members can safely express their views, but at the expense of identifying who supported the resulting policy. This lack of traceability—a record linking votes to identities—can in turn affect downstream decisions, such as the selection of future university leaders. In voting environments with social-image concerns, there is therefore a tension between truthful expression and vote traceability. This tension raises a natural question: is there a simple voting mechanism that preserves truthful expression while still allowing a degree of traceability of individual votes?

In this paper, we propose such a mechanism, which we call *threshold majority voting*. We first characterize its properties theoretically and then test its performance in an environment—a college campus—that has been at the center of recent debates about free expression in the United States. Threshold majority voting works as follows: first, individuals cast a vote over a policy; second, each non-abstaining voter also chooses a *disclosure threshold* that governs whether and when her individual vote is publicly revealed. A person's vote is publicly disclosed together with her name if and only if the share of votes for the option she selected exceeds her chosen threshold; otherwise, her individual vote remains undisclosed.

To study the theoretical properties of threshold majority voting, we develop a simple framework that embeds social-image concerns into a voting environment. Under anonymous majority voting, participation is high and voting is truthful, but nothing is learned about individual behavior. Under public majority voting, in contrast, one learns who supported which option, but both participation and vote choices are distorted. Threshold majority voting can resolve this tension by restoring the high-participation, truthful-voting equilibrium of anonymous majority voting while still revealing some information about individual votes. The mechanism operates through three distinct channels: it offers privacy to privacy-conscious individuals, it allows supporters of the controversial option to reveal their vote only when they have "safety in numbers," and it creates an epistemic force whereby one's vote is revealed precisely when it is more likely to match the underlying state of the world.

In order to assess whether the threshold mechanism delivers its desirable properties in practice, we conducted an experiment on a U.S. college campus. This setting is particularly fitting for two reasons. First, because self-censorship driven by social-image concerns is well documented on U.S. college campuses, establishing the preconditions for our mechanism to be valuable (Braghieri 2024; Ho and Huang 2024). Second, because college campuses have become flashpoints in national debates over free expression, with universities facing increased scrutiny from politicians, donors, and regulators.

Our experiment, conducted at the University of California, Berkeley, centers on a contentious policy question: whether transgender women should be allowed to compete in women's collegiate sports. Student participants are informed that they will be asked to cast a vote on this question and that the aggregate results will be shared with the UC Berkeley Chancellor, potentially informing future university decisions. Participants are then randomized into one of three treatments: (i) a *Private* treatment, in which individual votes are completely anonymous; (ii) a *Public* treatment, in which individual name-vote pairs are made public; and (iii) a *Threshold* treatment, in which a voter's name is revealed together with her vote only if her disclosure threshold is met.

Our first set of experimental results highlights the tension between truthful expression and vote traceability that motivates the paper. Moving from anonymous to public majority voting increases abstention rates by 70% and substantially reduces support for the socially controversial option, shifting the vote margin toward the socially uncontroversial one.[1] The drop in support for the controversial option is so large as to flip the collective decision: under public majority voting, the majority of voting participants supports the inclusion of transgender women in women's collegiate sports; under anonymous majority voting, the

[1] In our context, the socially controversial option is voting against transgender women participating in women's collegiate sports. See Section 1 for details.

majority of voting participants supports the opposite policy. Our experimental setting is therefore one in which truthful expression and vote traceability are in direct conflict, which is the precondition for threshold majority voting to be potentially valuable.

Our main set of experimental results examines the performance of the threshold mechanism. In line with the theoretical model, threshold majority voting mitigates the conflict between truthful expression and traceability. In particular, voting behavior under threshold majority voting is statistically indistinguishable from behavior under anonymous majority voting. At the same time, a sizable fraction (one third) of participants in the Threshold treatment have their individual vote revealed. Importantly, the public record of disclosed votes includes not only participants who supported the uncontroversial option but also those who supported the controversial one. Overall, therefore, threshold majority voting is able to elicit public expression of controversial opinions without sacrificing truthful voting.

From a policy perspective, threshold majority voting is potentially valuable when two conditions hold: (i) individual votes are relevant for downstream evaluation of or selection among the individuals casting them; and (ii) the issue is sufficiently socially charged that public observability creates social-image distortions in participation or vote choice. These conditions are common in small- and medium-sized electorates such as governance bodies and committees (universities, nonprofits, and public agencies) voting on socially sensitive or identity-related policies; professional associations and licensing boards adjudicating ethics rules, disciplinary actions, or contested standards; and corporate boards or workplace councils making decisions that may trigger reputational backlash (e.g., DEI, labor, or political-speech policies). In such environments, threshold majority voting provides a simple, implementable alternative to both anonymous and public voting: it preserves the preference- and information-aggregation properties of anonymous majority rule while still producing a public record of some individual votes, thereby enabling the partial assignment of responsibility for the implemented choice and facilitating downstream coordination. More broadly, threshold disclosure could help institutions elicit and aggregate sensitive information—such as reports of sexual harassment, discrimination, retaliation, or safety violations—by protecting isolated complainants and early whistleblowers from being singled out while still ensuring that, once concern is sufficiently widespread, some reports become traceable.

From an implementation perspective, the mechanism can be run without relying on a trusted mechanism designer (Akbarpour and Li 2020). For example, ballots and disclosure thresholds can be submitted as cryptographic commitments (or encrypted votes), and the disclosure rule can be enforced automatically once aggregate support crosses the relevant thresholds, with public verifiability that disclosures occurred exactly as specified. Implemented via a smart contract on a blockchain (or permissioned ledger), this approach can make both aggregation and conditional disclosure auditable, while ensuring that no administrator can selectively learn, reveal, suppress, or alter individual votes.

The paper contributes to several strands of literature. First, it adds to the growing body of work on social image, social desirability, and public expression (Bursztyn and Jensen 2017). Existing research shows that social-image concerns can affect behavior in a wide range of domains, including the expression of political attitudes (Kuran 1997; Bursztyn et al. 2023; Bursztyn, Egorov, and Fiorin 2020), voting (Kuran 1987; Dellavigna et al. 2017; Funk 2010; Gerber, Green, and Larimer 2008), political participation (Perez-Truglia and Cruces 2017), prosocial behavior (Bénabou and Tirole 2006), educational effort (Bursztyn and Jensen 2015; Bursztyn, Egorov, and Jensen 2019), and labor force participation (Bursztyn, González, and Yanagizawa-Drott 2020). On college campuses, recent work documents sizable wedges between privately held and publicly portrayed attitudes on sensitive political issues and shows that these wedges reduce the informativeness of public statements (Braghieri 2024; Ho and Huang 2024). We build on this work by shifting the focus from measuring distortions to designing institutions that mitigate them. In particular, we introduce a voting mechanism that addresses the core tension between truthful expression and vote traceability. In both the model and the experiment, threshold majority voting delivers the same voting outcomes as anonymous majority voting while revealing partial information about individual votes. It therefore helps elicit the expression of socially stigmatized views without distorting voting behavior.

Second, we contribute to the literature on voting, information aggregation, and institutional design. A large theoretical literature studies how and when majority rule aggregates private information (Austen-Smith and Banks 1996; Feddersen and Pesendorfer 1997), how abstention affects outcomes (Feddersen and Pesendorfer 1996), and how transparency shapes incentives in committees and legislatures (Levy 2007; Prat 2005; Visser and Swank 2007). A recurring theme in this work is that transparency can distort behavior: public voting may induce herding, pandering, or reputational posturing that undermines information aggregation (Maskin and Tirole 2004). Much of this literature treats the choice between secret and public ballots as discrete (Mattozzi and Nakaguma 2023). We instead study a mechanism that nests both extremes and gives individuals a choice over whether, and under what conditions, their vote is disclosed. Our results show that partial transparency implemented via threshold majority voting can deliver the same aggregate behavior as fully anonymous voting while providing some of the traceability benefits of public roll-call votes. This complements theoretical work on optimal transparency and privacy in environments with social-image concerns (Ali and Bénabou 2020; Levy 2007) and provides experimental evidence on the performance of this alternative voting procedure.

Lastly, our analysis speaks to how formal institutions interact with informal institutions, and how the former can be designed to mitigate the shortcomings of the latter (Acemoglu and Jackson 2017). A growing theoretical literature studies how laws and formal rules coexist with norms, values, and conventions, and how their effects depend on this interaction rather than on either set of institutions in isolation (Helmke and Levitsky 2004; Acemoglu and Jackson 2017; Bénabou and Tirole 2025). We study a setting in which informal pressures generate systematic distortions in public voting, and we show that a simple change in the formal decision rule can dampen these distortions without requiring a change in underlying norms. In this sense, the paper illustrates how formal institutional design can be used to offset the unintended consequences of informal constraints. More broadly, our results point to a strategy for institutional design in environments with strong informal pressures: rather than attempting to eliminate those pressures, formal rules can be structured so that their interaction with informal institutions still achieves desirable properties.

The rest of the paper is organized as follows. Section 1 provides some useful context for our experimental investigation. Section 2 presents the motivating framework that illustrates the properties of threshold majority voting. Section 3 describes our experimental design. Section 4 discusses the results. Section 5 concludes.

1 Background: College Campuses and Policy Context

College campuses are a natural setting in which to study how public observability and social image shape political expression. Universities train future elites, play a central role in forming political attitudes and norms of democratic discourse, and routinely ask students to take positions on controversial issues in front of their peers.

In recent years, campus speech has also become a focus of national political conflict. Recent administrations have sought to reshape higher education through executive orders that tie federal research funds to campus free-speech policies, restrict diversity, equity, and inclusion (DEI) initiatives, and threaten funding cuts or legal action against universities over protest activity and policies on gender and race. At the same time, debates over "political correctness," DEI, and the appropriate boundaries of campus discourse have intensified in the public sphere. Recent survey and experimental work shows that students' public statements on politically sensitive topics are often systematically distorted by social desirability concerns, leading to gaps between privately held beliefs and what students are willing to say in public (Braghieri 2024; Ho and Huang 2024). These features make college campuses a particularly appropriate environment in which to evaluate the performance of threshold majority voting.

We study these issues in the context of the University of California, Berkeley (UC Berkeley), a large public university that is widely perceived as politically liberal and that figures prominently in contemporary debates over campus speech. In our experimental sample, the ideological distribution among students is highly skewed: the ratio of self-identified liberals to conservatives is 10.2:1 (75.5% liberal vs. 7.4% conservative). External indicators are consistent with a strained speech climate. In the Foundation for Individual Rights and Expression (FIRE) College Free Speech Rankings (Stevens 2025), UC Berkeley receives an overall speech climate grade of F and is ranked 217th out of 257 institutions, with relatively low overall scores and weak performance on measures of comfort expressing ideas. Consistent with these indicators, 45.5% of students report self-censoring on campus at least once or twice a month (see Online Appendix Figure A1).

This combination of ideological imbalance and perceived speech constraints is central to our research question. When one political position is perceived as dominant in the local environment, individuals whose views depart from the perceived norm may face stronger stigma from expressing those views publicly, and even those whose views align with the local majority may feel pressure to adopt more extreme public positions. In such settings, public observability of individual political actions can both distort participation (who chooses to speak or vote) and bias expression (what people are willing to say), making public signals less informative about underlying preferences or beliefs than anonymous ones. The campus environment at UC Berkeley therefore offers a natural laboratory for studying how alternative voting mechanisms interact with these forces.

Policy proposal. Within this broader environment, our experiment centers on a concrete policy question: whether UC Berkeley should allow transgender women to compete in women's collegiate sports. We chose this issue for three reasons. First, because the topic is highly salient and politically relevant on college campuses. Questions about the participation of transgender athletes in sex-segregated sports have become a focal point of student politics, media coverage, and university governance, and they speak directly to broader debates about equity, inclusion, and fairness. Second, because prior survey evidence at UC Berkeley (Ho and Huang 2024) documents strong social desirability pressures around this specific issue: public expression diverges sharply from private beliefs, and stated positions are sensitive to perceptions of the majority opinion. Third, because the issue involves genuine disagreement despite these pressures. As shown in our experiment, a non-trivial share of students privately hold views that diverge from the view they perceived to be more socially appropriate on campus, creating precisely the conditions under which a mechanism like threshold majority voting can improve truthful expression relative to fully public voting.

The policy question is also embedded in a contentious legal and regulatory landscape. 29 out of 50 U.S. states have enacted laws or regulations restricting the participation of transgender athletes in school sports, typically requiring that student-athletes compete on teams corresponding to their sex assigned at birth rather than their gender identity (Online Appendix Figure A2). These laws vary in scope across states, but collectively they create a patchwork of eligibility rules that affect transgender youth and collegiate athletes. At the national level, the National Collegiate Athletic Association (NCAA) recently overhauled its participation policy for transgender student-athletes. As of February 2025, competition in NCAA women's sports is limited to athletes who were assigned female at birth, reversing earlier sport-by-sport guidelines and aligning eligibility rules more closely with recent federal executive guidance (NCAA 2025; Executive Order 14220, 2025). These developments have made the inclusion of transgender athletes in women's sports a central point of legal, political, and cultural conflict in the United States.

The next section develops a simple theoretical framework that formalizes the trade-offs between anonymous, public, and threshold majority voting and generates testable predictions for our experimental setting.

2 Motivating Framework

We develop a simple theoretical framework to motivate the empirical analysis. The model clarifies how public observability shapes both participation and vote choice, and how a threshold mechanism can mitigate these effects.

In environments with strong social-image concerns, public expression generates two well-documented distortions relative to private expression (Braghieri 2024; Bursztyn and Jensen 2017; Ho and Huang 2024). First, it discourages some individuals from voicing their opinions at all (extensive margin). Second, among those who do express their views publicly, it shifts stated opinions toward the socially uncontroversial option and away from truthful expression (intensive margin). In a voting context, this translates to higher abstention rates and a higher share of votes for the socially uncontroversial option under public majority rule than under anonymous majority rule.

The threshold mechanism we introduce allows each voter to decide when her vote will be publicly revealed: her vote becomes observable only if the share of votes for the option she picked exceeds a privately chosen threshold. This design aims to alleviate the costs of public voting through three distinct channels.

Privacy channel. Threshold majority voting allows privacy-conscious individuals to participate without their vote appearing in the public record, removing a publicity cost that might otherwise deter them from voting.

Safety-in-numbers channel. If the stigma from supporting the socially controversial option declines in the fraction of people who vote for it, threshold majority voting allows voters to reveal their votes only when enough others voted the same way; that is, only when the stigma cost is sufficiently low.

Epistemic channel. With imperfect private signals about an underlying state of the world, truthful voting implies that observing many others vote the same way provides evidence that one's signal was likely correct. Threshold majority voting allows individuals to reveal their vote precisely when their controversial choice is more likely to match the state of the world.

The rest of this section formalizes the observations above in a stylized environment.

2.1 Setup

There are two states of the world, $\omega \in \{0, 1\}$, each realized with probability 1/2. A continuum of agents of total mass 1, indexed by $i \in [0, 1]$, is called to vote on a binary policy concerning a controversial issue. Each agent chooses an action $a_i \in \{0, 1, \tilde{a}\}$, where $a_i \in \{0, 1\}$ represents a vote for one of the two policies and $\tilde{a}$ denotes abstention. We assume that the socially optimal policy in state ω is $a = \omega$. We also assume that option $a = 0$ is socially uncontroversial, whereas option $a = 1$ is socially controversial. We let $\bar{a}$ denote the policy implemented by the voting mechanism.

Signals and types. Each agent i receives a private signal $s_i \in \{0, 1\}$ about the realized state of the world ω. Signals are i.i.d. across agents and diagnostic of the state. In particular, we let $\Pr(s_i = \omega \mid \omega) = \lambda$ and $\Pr(s_i = 1 - \omega \mid \omega) = 1 - \lambda$, with $\lambda \in (1/2, 1)$. Because the electorate is a continuum, if all agents vote according to the signal they observe, the implemented policy matches the realized state of the world almost surely.

Besides differing in the signal they observe, agents also differ in social-image concerns, privacy concerns, and participation costs. We denote agent i's type by $\boldsymbol{x}_i = (s_i, \eta_i, \pi_i, c_i)$, where each component is interpreted as follows. $s_i \in \{0, 1\}$ is the agent's private signal about the state of the world. $\eta_i \in \{\eta_L, \eta_H\}$ captures the idiosyncratic extent to which the agent cares about the stigma arising from being perceived as supporting the socially controversial option $a = 1$. $\pi_i \in \{\pi_L, \pi_H\}$ is the idiosyncratic extent to which the agent cares about having her individual vote publicly revealed. $c_i \in \{c_L, c_H\}$ captures agent i's cost of casting a vote rather than abstaining. We assume $\eta_L < \eta_H$, $\pi_L < \pi_H$, and $c_L < c_H$, and, to simplify computations, impose the normalization $\eta_L = \pi_L = c_L = 0$.

We assume that (η_i, π_i, c_i) is independent of (s_i, ω) and that η_i, π_i and c_i are mutually independent. For $x \in \{\eta_L, \eta_H, \pi_L, \pi_H, c_L, c_H\}$, we denote the prior probability of each realization by p_x. Thus, conditional on state ω, the probability that agent i is of type $(s_i, \eta_i, \pi_i, c_i)$ is $\Pr(s_i, \eta_i, \pi_i, c_i|\omega) = \Pr(s_i|\omega)p_{\eta_i}p_{\pi_i}p_{c_i}$. We assume all types have strictly positive mass.

External audience and information. To capture reputational forces, we assume that an external audience forms beliefs about each agent's voting behavior and that these beliefs enter the agent's payoff. The audience shares the common prior over ω and knows the mechanism, the type distributions, the signal structure, and the equilibrium strategy profile. However, it has no individual-level information about any agent beyond knowing that all agents are ex ante identical draws from the type distribution and beyond whatever individual-level information is revealed by the voting mechanism.

Each mechanism generates a public signal, denoted by Ψ, which summarizes the information that is observable to the audience (ranging from aggregate vote shares to individual vote labels, depending on the mechanism). After observing the public signal Ψ and the realized state, the audience forms posterior beliefs about each agent's action. We denote these beliefs by $P_j(a_i = \cdot \mid \Psi, \omega)$: audience j's Bayesian posterior over agent i's vote.[2]

Voting mechanisms. We compare three mechanisms—anonymous majority voting, public majority voting, and threshold majority voting—that all implement majority rule among non-abstainers but differ in the observability of individual votes. Each mechanism generates a different public signal, denoted Ψ^{Pri}, Ψ^{Pub}, and Ψ^{Thr}.

Anonymous Majority Voting. The public signal under anonymous (private) majority voting, Ψ^{Pri}, discloses only aggregate outcomes: the fraction of abstentions and the vote

[2] The subscript j does not index different audiences. There is a single audience; we employ the subscript j only to emphasize that these posterior beliefs are taken from the audience's perspective, which is based on an information set that is different from the agents'.

shares for $a = 0$ and $a = 1$. Individual participation decisions and individual votes are never revealed.

Public Majority Voting. The public signal under public majority voting, Ψ^{Pub}, discloses, for every agent, whether she abstained and, if she voted, which option she selected. Thus all individual votes are publicly observable and so are individual abstentions.

Threshold Majority Voting. Under threshold majority voting, each agent who does not abstain also chooses a disclosure threshold $t_i \in [0, 1]$. Her individual vote is revealed if and only if the realized share of votes for her chosen option is at least t_i.[3] Thus, the public signal under threshold majority voting, Ψ^{Thr}, discloses aggregate vote shares as well as the identities and votes of agents whose thresholds are met; all remaining agents—including abstainers—appear as "undisclosed." Individuals' thresholds choices are not publicly disclosed in Ψ^{Thr}.

Timing. The timing of the game is as follows: first, nature draws the state ω and, conditional on ω, each agent's type $\boldsymbol{x}_i = (s_i, \eta_i, \pi_i, c_i)$. Second, each agent privately observes her own type $\boldsymbol{x}_i$. Third, agents simultaneously choose an action $a_i \in \{0, 1, \tilde{a}\}$. Under the Threshold mechanism, each agent who casts a vote ($a_i \neq \tilde{a}$) also chooses a disclosure threshold $t_i \in [0, 1]$. Fourth, votes are aggregated and the policy $\bar{a}$ is implemented by majority rule among non-abstainers (with ties resolved by fair randomization). Fifth, a public signal Ψ is generated according to the mechanism's disclosure rule. Sixth, the realized state of the world ω is revealed. Seventh, the external audience observes (ω, Ψ) and forms beliefs about each agent's action. Eighth, payoffs are realized.

Preferences. Agents' payoffs depend on six forces generally considered relevant to voting behavior: (i) instrumental concerns (Downs 1957), (ii) expressive benefits (Brennan and Lomasky 1993), (iii) reputation-for-accuracy concerns (Levy 2007), (iv) social stigma (Kuran 1987), (v) privacy concerns (Gerber et al. 2013), and (vi) participation costs (Downs 1957). We describe each component in turn and then present the payoff function that includes all of them. Throughout, we let $\mathbb{E}_i[\cdot]$ denote agent i's ex ante expectation.

[3] To give agents a unilateral "never reveal" option, we assume that setting $t_i = 1$ guarantees that an agent's vote is never individually disclosed, regardless of the realized vote shares.

Instrumental concerns. Agents derive instrumental benefits from the implemented policy $\bar{a}$ if it matches the realized state of the world. Letting $\beta \geq 0$ measure the strength of this motive, the instrumental benefits are $\beta\,\mathbf{1}\{\bar{a} = \omega\}$.

In a continuum model, each individual agent is non-pivotal: conditional on her type, her action has no effect on the probability that $\bar{a} = \omega$. As a result, the term $\beta\,\mathbb{E}_i[\mathbf{1}\{\bar{a} = \omega\}]$ does not affect marginal incentives and will be treated as a constant when analyzing the agent's choice problem. We include it in the payoff function for completeness.

Expressive benefits. Agents derive a direct benefit from voting in line with their private signal. Let $\phi > 0$ denote the strength of this expressive motive. Agent i receives ϕ if she votes according to her signal and 0 otherwise. The expressive benefit is thus $\phi\,\mathbf{1}\{a_i = s_i\}$.

Reputation-for-accuracy concerns. Agents care about being perceived as having voted in favor of the policy that matches the realized state of the world. After observing (Ψ, ω), the external audience forms a posterior belief about each agent's action. The audience's posterior that agent i supported the policy that matches the realized state of the world is $P_j(a_i = \omega \mid \Psi, \omega)$. Agent i's reputation-for-accuracy payoff is proportional to her expectation of the probability that the audience assigns to her having matched the state: $\kappa\,\mathbb{E}_i[\,P_j(a_i = \omega \mid \Psi, \omega)\,]$, where $\kappa > 0$.

Social stigma. Agents suffer a stigma cost from being perceived as having voted for the socially controversial option $a = 1$. This cost has two components: an idiosyncratic intensity $\eta_i \in \{\eta_L, \eta_H\}$ and a common "safety-in-numbers" term that penalizes voting for the controversial option less heavily when the vote share for that option is larger.

The safety in numbers term is captured by

$$f(\mu(\Psi)) = \begin{cases} \min\{\frac{1}{\mu(\Psi)}, M\} & \text{if } \mu(\Psi) > 0, \\ M & \text{if } \mu(\Psi) = 0 \end{cases}$$

where $M > 0$ is a constant and $\mu(\Psi) \in [0, 1]$ denotes the realized share of votes for $a = 1$ among non-abstainers, as described by the public signal Ψ.[4] We let $P_j(a_i = 1 \mid \Psi, \omega)$ denote

[4] The cap on $f(\cdot)$ serves two purposes. First, it ensures that the stigma term is well defined even when

the audience's posterior, after observing (Ψ, ω), that agent i supported the controversial option. Agent i's expected stigma cost is thus $\eta_i \, \mathbb{E}_i[f(\mu(\Psi)) \, P_j(a_i = 1 \mid \Psi, \omega)]$.

Privacy concerns. Each non-abstaining agent i pays a privacy cost π_i whenever her individual vote appears in the public record Ψ. Letting $\mathbf{1}\{a_i \in \{0,1\} \text{ and } \Psi \text{ reveals } a_i\}$ denote whether agent i's vote is revealed by the public record, the expected privacy costs are $\pi_i \, \mathbb{E}_i[\mathbf{1}\{a_i \in \{0,1\} \text{ and } \Psi \text{ reveals } a_i\}]$.

Under anonymous voting, privacy costs are always zero: individual votes are never revealed. Under public voting, abstainers do not pay the privacy cost, whereas non-abstainers do. Under threshold voting, abstainers do not pay the privacy cost; non-abstainers pay it if and only if the realized support for the option they voted for meets or exceeds the threshold t_i they set.

Participation costs. Casting a vote (for either $a = 0$ or $a = 1$) requires paying a participation cost c_i. Abstaining avoids this cost. The participation cost term is $c_i \, \mathbf{1}\{a_i \neq \tilde{a}\}$.

Putting these components together, agent i's expected utility from choosing action $a_i \in \{0, 1, \tilde{a}\}$ (and threshold $t_i \in [0,1]$ under threshold majority voting) is

$$\begin{aligned} u_i(\boldsymbol{x}_i, a_i, t_i) = \; & \underbrace{\beta \, \mathbb{E}_i[\mathbf{1}\{\bar{a} = \omega\}]}_{\text{instrumental concerns}} + \underbrace{\phi \, \mathbf{1}\{a_i = s_i\}}_{\text{expressive benefits}} + \underbrace{\kappa \, \mathbb{E}_i[P_j(a_i = \omega \mid \Psi, \omega)]}_{\text{reputation-for-accuracy concerns}} \\ & - \underbrace{\eta_i \, \mathbb{E}_i[f(\mu(\Psi)) \, P_j(a_i = 1 \mid \Psi, \omega)]}_{\text{social stigma}} \\ & - \underbrace{\pi_i \, \mathbb{E}_i[\mathbf{1}\{a_i \in \{0,1\} \text{ and } \Psi \text{ reveals } a_i\}]}_{\text{privacy concerns}} - \underbrace{c_i \, \mathbf{1}\{a_i \neq \tilde{a}\}}_{\text{participation costs}} \end{aligned} \tag{1}$$

Under anonymous and public majority voting, thresholds are not chosen and we write $u_i(\boldsymbol{x}_i, a_i)$ with the same components.

We assume that, whenever an agent is indifferent across multiple thresholds, she chooses the largest such threshold. If the indifference set does not admit a maximum, we allow her to

$\mu(\Psi) = 0$, in which case $1/\mu(\Psi)$ is not defined. Second, it guarantees that the payoff function remains bounded. In the equilibrium analysis, the share $\mu(\Psi)$ is strictly positive on the equilibrium path, and we choose M large enough that the cap never binds in equilibrium. Thus, M is included for definitional rigor, but it does not play a substantive role in the analysis.

pick an arbitrary threshold in that set.[5]

Equilibrium notion. Given the payoff functions and agents' private information encoded in their types, the interaction defines a Bayesian game with a public signal. We focus on (weak) Perfect Bayesian Equilibria, referred to henceforth simply as equilibria.

An equilibrium consists of: (i) a strategy profile σ that maps each type $\boldsymbol{x}_i$ into a (possibly mixed) action $a_i \in \{0, 1, \tilde{a}\}$. Under threshold majority voting, σ also specifies a (possibly mixed) disclosure threshold $t_i \in [0, 1]$ whenever $a_i \neq \tilde{a}$. (ii) A belief system that, for each public signal Ψ and state ω, assigns posterior probabilities $P_j(a_i = \cdot \mid \Psi, \omega)$ to every agent's action.

Strategies and beliefs must satisfy two conditions. First, given the belief system and the strategies of other agents, each type's strategy maximizes that type's expected payoff. Second, beliefs are consistent with Bayes' rule at all public signals Ψ, given the mechanism, the type distribution, and the strategy profile.[6]

2.2 Main Results

We begin our analysis by formalizing the main tension between anonymous and public majority voting that motivates the introduction of the threshold mechanism. Anonymous majority voting preserves high participation, induces truthful voting, and selects the policy that matches the state of the world almost surely, but it reveals nothing about individual votes. Public majority voting, in contrast, increases transparency by revealing who supported which option, but it distorts both participation and vote choices through social-image concerns, thereby reducing decision quality. The following proposition formalizes this benchmark comparison in our environment. The proofs of all the propositions are relegated to Appendix B.

[5] Under the equilibrium strategies characterized in the analysis of threshold majority voting, indifference only arises on closed intervals that have a well-defined maximum.

[6] Off-path beliefs arise only in a limited sense in our environment because the public signal Ψ is, by construction, a sufficient statistic for all payoff-relevant public information. Any unilateral deviation that changes the realized public record from Ψ to Ψ' is publicly described by Ψ' itself; the audience's posterior is therefore the Bayesian posterior conditional on (Ψ', ω) under the mechanism and the strategy profile. Deviations that do not change Ψ are observationally irrelevant to the audience.

Proposition 1. *There exist $\bar{p}_{c_H}, \bar{c}, \bar{\pi}, \bar{\phi}, \underline{\eta}, \bar{\eta} > 0$ with $\underline{\eta} < \bar{\eta}$ such that, if $p_{c_H} > \bar{p}_{c_H}$, $c_H > \bar{c}$, $\pi_H > \bar{\pi}$, $\phi > \bar{\phi}$, and $\eta_H \in (\underline{\eta}, \bar{\eta})$, the following holds. There exists an equilibrium under public majority voting with the following features:*

- ***Participation margin.*** *The abstention rate is strictly higher than in the unique equilibrium of anonymous majority voting.*

- ***Expression of the socially controversial view.*** *The fraction (out of the entire population) of agents voting for the socially controversial option $a = 1$ is strictly lower than in the unique equilibrium of anonymous majority voting.*

- ***Vote margins.*** *Support for the socially non-controversial option $a = 0$ among non-abstainers is strictly higher than in the unique equilibrium of anonymous majority voting. Moreover, in the unique equilibrium of anonymous majority voting, all agents who choose to participate vote truthfully ($a_i = s_i$). As a result, the implemented policy coincides with the realized state of the world almost surely, $\bar{a} = \omega$. In contrast, if the population share of high-stigma types is sufficiently large, the considered equilibrium of public majority voting always implements the socially non-controversial policy, $\bar{a} = 0$, regardless of the realized state of the world.*

The proposition relies on several restrictions on the primitives of the model; we discuss these restrictions in detail in Section 2.3.

We now turn to the main proposition of the paper, which shows that threshold majority voting reconciles the tension highlighted in Proposition 1. Specifically, by giving voters control over when their vote becomes visible, it protects truthful expression while still allowing for a meaningful degree of disclosure.

Proposition 2. *Under the same parameter restrictions as in Proposition 1, there exists a unique equilibrium under threshold majority voting with the following features:*

Voting behavior.

- ***Participation margin.*** *The abstention rate coincides with that in the unique equilibrium of anonymous majority voting.*

- ***Expression of the socially controversial view.*** *The fraction (out of the entire population) of agents voting for the socially controversial option* $a = 1$ *coincides with that in the unique equilibrium of anonymous majority voting.*

- ***Vote margins.*** *Support for the socially non-controversial option* $a = 0$ *among non-abstainers coincides with that in the unique equilibrium of anonymous majority voting. As a result, the implemented policy coincides with the realized state of the world almost surely,* $\bar{a} = \omega$*.*

Vote traceability and disclosure.

- ***Disclosure under threshold majority voting.*** *A strictly positive fraction of voters for each option,* $a = 0$ *and* $a = 1$*, have their votes publicly revealed.*

- ***Threshold choices.*** *The distribution of disclosure thresholds chosen by* $a = 1$ *voters first-order stochastically dominates the distribution chosen by* $a = 0$ *voters. Thus, individuals who voted for the socially controversial option require a larger fraction of like-minded voters in order to be willing to disclose their vote.*

Taken together, the two propositions show how the three mechanisms address the core tension between truthful expression and vote traceability. Anonymous majority voting performs well on decision quality—many agents participate and all participating agents vote truthfully—but it provides no information about individual behavior and thus offers little scope for assigning partial responsibility for the implemented decision. Public majority voting moves too far in the opposite direction: it maximizes traceability by revealing every vote but, in doing so, it induces conformity and additional abstentions, impairing the aggregation of information. Threshold majority voting offers a middle ground. By giving individuals control over when their vote becomes visible, it preserves the information–aggregation properties of anonymous majority voting while providing traceability through partial disclosure of who supported which option.

Proposition 2 also clarifies how the three channels that make threshold majority voting intuitively appealing operate in equilibrium through the payoff components.[7]

[7] See Online Appendix B.3 for a description of where each of the three channels appears in the proof of Proposition 2.

Privacy channel: because the privacy cost π_i is incurred only when the mechanism records an agent's name–vote pair in Ψ, threshold majority voting allows privacy-conscious voters (π_H) to participate while choosing thresholds that keep them "undisclosed" with probability one, thereby removing the publicity cost that deters them from participating under public voting.

Safety-in-numbers channel: the stigma term penalizes being perceived as having supported the controversial option more heavily when its vote share $\mu(\Psi)$ is small. Threshold majority voting allows voters for the socially controversial option to disclose their vote only when the vote share for their preferred option is sufficiently high and, thus, the stigma cost is sufficiently low.

Epistemic channel: the reputation-for-accuracy payoff enters through $\kappa\,\mathbb{E}_i[P_j(a_i = \omega \mid \Psi, \omega)]$, which rewards being perceived as having matched the realized state of the world. Threshold majority voting enables agents to reveal their vote when their controversial choice is ex post more likely to coincide with the realized state of the world.

2.3 Discussion of Modeling Choices

In this section, we briefly discuss some of our modeling choices and relate them to the design of our experiment.

Restrictions on primitives in Propositions 1 and 2. Propositions 1 and 2 impose restrictions on the model's primitives. We discuss each in turn.

We impose a restriction on ϕ to generate a meaningful trade-off between truthful expression and vote traceability. If reputation-for-accuracy benefits were too large relative to ϕ, then publicly choosing the option an agent deems less likely to be correct could outweigh the direct gain from truthful expression. In that case, an agent might publicly support an option she privately believes is less likely to be correct purely to gain "reputation-for-accuracy points" in the low-probability event that her private signal turns out to be wrong. We rule out this counterintuitive behavior by requiring ϕ to be sufficiently large.

The restriction on π_H is intended to capture the empirical regularity, documented in prior work (Ho and Huang 2024), that publicity reduces expression for both individuals who hold the socially controversial view and those who hold the socially uncontroversial view, albeit potentially to different extents.

The lower bound on η_H ensures that public majority voting actually distorts behavior. If η_H were too small, social-image concerns would be negligible and public observability would not meaningfully affect voting relative to anonymous majority voting.

The lower bounds on c_H and p_{c_H} and the upper bound on η_H, while not crucial for Proposition 1, are essential for threshold majority voting to exhibit the desirable properties outlined in Proposition 2.

The restrictions on c_H and p_{c_H} ensure the existence of a sufficiently large mass of agents who abstain for non-strategic reasons (those with $c_i = c_H$). Under threshold majority voting, this pool provides "social cover" for individuals who support the controversial option and choose high disclosure thresholds. The intuition is simple: since some agents genuinely abstain for idiosyncratic reasons—being sick, traveling, being insufficiently interested in or informed about the topic, etc.—appearing as "undisclosed" does not automatically imply that one voted for the socially controversial option. We view this as a realistic feature of many voting environments, where participation rates are rarely 100%. It also matches our experimental environment, where roughly 20% of students abstain even under anonymous majority voting.[8]

A second crucial restriction for threshold majority voting to deliver its desirable properties is that $\eta_H < \bar{\eta}$. If η_H were too large, avoiding social stigma would dominate all other motives, making truthful voting unattractive even under threshold majority voting. Our assumptions therefore place η_H in an intermediate region $(\underline{\eta}, \bar{\eta})$ in which (a) public observability distorts behavior, and (b) threshold majority voting can mitigate those distortions.

Latent state of the world. We assume a latent state of the world $\omega \in \{0, 1\}$, informative private signals s_i, and agents who have some desire to be perceived as having supported the policy that matches the realized state of the world. This structure serves two purposes. First, it allows us to study the information-aggregation properties of the three voting mechanisms,

[8] If one were concerned that the abstaining pool might be too small, one could create additional social cover by randomly selecting a subset of individuals to cast an anonymous ballot. This way, remaining "undisclosed" would not uniquely signal a controversial preference.

which is a central theme in the voting literature since Condorcet's jury theorem (Condorcet 1785). Second, it makes the epistemic channel of the threshold mechanism transparent: because equilibrium vote shares are state-dependent, observing that "my side is large" is informative about the accuracy of one's private signal.

Importantly, the existence of a latent state of the world is not essential for the comparisons we draw in this section. In particular, one can dispense with ω altogether and, instead, assume that agents receive utility from being perceived as having voted in accordance with their own private views. Under this alternative interpretation, the audience uses the public signal Ψ to form posteriors about whether an agent voted honestly; public observability can still generate abstention and conformity through social-image concerns; and the threshold mechanism can still mitigate these distortions by allowing conditional disclosure. Moreover, the formal analysis and comparative statics are essentially unchanged: replacing the "matching the state" object $P_j(a_i = \omega \mid \Psi, \omega)$ with an audience posterior about honest voting yields the same structure of incentives and the same qualitative predictions across mechanisms.

In our experiment, the policy question is normative rather than factual, so this "reputation for honesty" interpretation is particularly natural. Under this interpretation, revealing individual votes can have dynamic value by revealing information about who holds which view in the population, which can facilitate future coordination within committees or organizations (e.g., by helping actors anticipate who will support related actions or coalitions). We nonetheless prefer casting the baseline model with a latent state space because it makes the epistemic implications of the three voting mechanisms especially transparent.

2.4 Empirical Hypotheses

Propositions 1 and 2 suggest the following empirical hypotheses.

- H1: The abstention rate is equal under anonymous and threshold majority voting, and strictly lower than under public majority voting.

- H2: The share expressing the socially controversial view (among all potential voters) is equal under anonymous and threshold majority voting, and strictly higher than under public majority voting.

- H3: The vote share for the socially uncontroversial option (among non-abstainers) is equal under anonymous and threshold majority voting, and strictly lower than under public majority voting.

- H4: A positive fraction of voters for both the socially uncontroversial and the socially controversial option have their individual name-vote pairs publicly revealed under threshold majority voting.

- H5: The distribution of threshold choices among voters for the socially controversial option first-order stochastically dominates that among voters for the socially uncontroversial option.

The next section describes the experimental implementation that allows us to test these predictions in a controlled yet policy-relevant environment.

3 Experimental Design

Our experiment implements the three voting mechanisms from Section 2—anonymous majority voting, public majority voting, and threshold majority voting—in the context of a college campus. Figure 1 summarizes the experimental flow by treatment; full materials, including survey screens and video transcripts, appear in Online Appendix D.

Recruitment. We conducted an online experiment with UC Berkeley undergraduates recruited through the campus experimental social science lab (XLab). Students received an email invitation and, upon accepting, were directed to an online survey platform, where they provided informed consent before proceeding.

Baseline instructions and policy proposal. After consenting to participate in the study, all participants watched a brief introductory video. The video explained the study in three steps: (i) participants would review a policy proposal, (ii) they would learn how their vote might be shared with other UC Berkeley students, and (iii) they would vote in favor of the proposal, vote against it, or abstain. The video emphasized that votes had real stakes: the

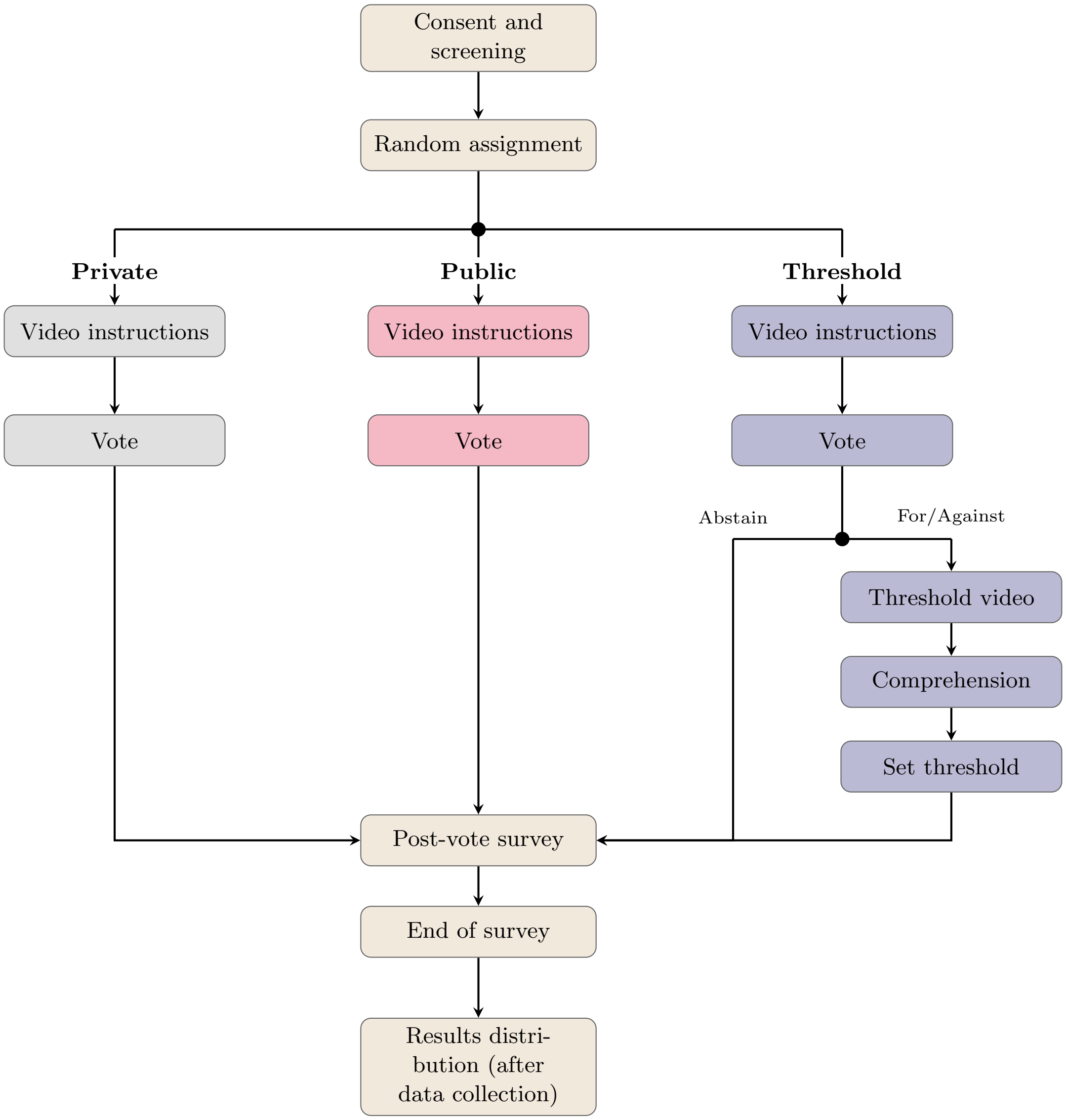


Figure 1: Experimental Flow by Treatment Condition

Note: This figure illustrates the experimental protocol by treatment condition. All participants completed the same initial stages: informed consent, video instructions explaining the study context and voting task, and comprehension checks. Participants were then randomly assigned to one of three treatments—Private (anonymous majority voting), Public (public majority voting), or Threshold (threshold majority voting)—and voted on the proposal (in favor, against, or abstain). *Private and Public treatments:* participants proceeded directly to the post-vote survey after voting. *Threshold treatment:* non-abstaining participants completed an additional stage that involved watching a second instructional video explaining the threshold mechanism, completing a practice round, passing threshold-specific comprehension checks, and setting their disclosure threshold. All participants then completed the post-vote survey. After data collection, results were distributed via email according to treatment-specific disclosure rules.

aggregate results of the vote (i.e., the percentage of students choosing each option) would be shared with the Chancellor's office and could inform future university policy.

Participants then saw the policy proposal: *"The university should allow transgender women to compete in women's collegiate sports."*

Random assignment and treatment-specific instructions. Participants were then randomly assigned with equal probability to one of three experimental treatments corresponding to the mechanisms in the model: *Private* (anonymous majority voting), *Public* (public majority voting), and *Threshold* (threshold majority voting).

Participants received treatment-specific video instructions explaining whether their vote would be linked to their name in the results sent to other students:

1. *Private.* The Private treatment implements anonymous majority voting. The video stated that the participant's vote would never be linked to their name or identity—it would remain completely anonymous. After the study, XLab staff would send out a spreadsheet listing the names of all participants in the treatment, but no votes would appear next to any name. In terms of the model, the public signal Ψ^{Pri} consists only of aggregate vote shares and a list of participants, with no individual vote labels.

2. *Public.* The Public treatment implements public majority voting. The video explained that the participant's vote would be visible to other UC Berkeley students in their treatment. After the study, XLab staff would send out a spreadsheet showing how everyone in the treatment voted: for each participant, the spreadsheet would list their name and, if they voted, their vote (for or against); abstainers would appear by name without a recorded vote. The public signal Ψ^{Pub} thus contains a name–vote record for all non-abstainers, and simply a name for all abstainers.

3. *Threshold.* The Threshold treatment implements threshold majority voting. The video informed participants that they would vote in two steps: first, they would cast their vote anonymously; second, they would decide whether to make their vote visible to other students by setting a disclosure threshold. The video introduced the threshold as the minimum fraction of *other non-abstaining participants* who must cast the *same vote* as

the participant for her vote to be shown next to her name in the results spreadsheet. It emphasized that, after the study, the names of all participants in the treatment would appear in the spreadsheet; if a participant's realized like-minded vote share reached or exceeded their chosen threshold, their vote would appear next to their name; otherwise their name would appear without a vote. The public signal Ψ^{Thr} thus contains aggregate vote shares and a partial name–vote record: some participants are labeled with their vote, and others appear as "undisclosed." This maps directly to the threshold majority voting mechanism in the model, where each non-abstaining agent chooses a threshold t_i and their name–vote pair is revealed if and only if the realized share of votes for their chosen option exceeds t_i.[9]

After viewing the treatment-specific video, participants answered comprehension questions to verify understanding of the experimental task. They were required to answer all questions correctly (with feedback after incorrect answers) before proceeding.

Voting decision. Once participants passed the comprehension checks, they moved to the voting screen. The interface displayed once again the policy proposal — *"The university should allow transgender women to compete in women's collegiate sports"* — and allowed participants to choose one of three options: vote in favor, vote against, or abstain.

The screen included a reminder of the participant's treatment-specific disclosure rule directly above the vote buttons. Participants in the Threshold treatment knew that, if they did not abstain, they would subsequently be asked to set a disclosure threshold.

Throughout the analysis, we treat voting *against* the proposal as the socially controversial option. This classification is motivated by prior evidence from the same population documenting strong perceived social norms in favor of transgender inclusion in sports (Ho and Huang 2024), and by our own post-vote survey measure, in which 79.9% of participants rated voting in favor the proposal as more socially acceptable than voting against (Online Appendix Figure A3).

[9] In line with the model, we gave students the option to unilaterally keep their vote anonymous by picking $t_i = 1$.

Threshold elicitation (Threshold treatment only). Participants in the Threshold treatment who did not abstain ($a_i \in \{0, 1\}$) proceeded to a threshold elicitation stage. This stage corresponds to the choice of disclosure threshold $t_i \in [0, 1]$ in the model and was implemented in several steps to ensure understanding:

1. *Additional instructional video.* Participants watched a second video explaining the threshold mechanism in detail, using simple numerical examples. The video defined the threshold as the minimum fraction of other non-abstaining participants who must choose the same option as the participant for her vote to be made public. It illustrated the extremes (0% threshold implies the vote is always public; 100% implies the vote is never public) and intermediate thresholds (e.g., 30%), and it showed cases in which the realized support was above or below the chosen threshold.

2. *Practice round and comprehension checks.* After viewing the instructional video, participants completed a brief practice exercise based on a hypothetical scenario about dining hall hours. They then answered three threshold-specific comprehension questions that tested their understanding of when a vote would be revealed or remain private under different threshold and vote-share combinations. Participants received immediate feedback and could make unlimited attempts. They were required to answer all questions correctly before proceeding.

3. *Binary-search elicitation.* Participants faced a sequence of questions of the form: "If at least X% of all voting students choose the same option as you, would you share your vote publicly?" After each response ("Yes, this threshold works for me" / "No, I need more students to agree with me (or I want to keep my vote private)"), the algorithm adjusted X up or down and posed a new question. This sequence converged to a narrow interval (within 5 percentage points), after which participants chose their exact disclosure threshold from a short list of values in that interval.

4. *Confirmation screen.* After selecting a threshold, participants viewed a confirmation screen summarizing their choice and its implications as a function of the realized vote share for the option they supported. They could opt to repeat the elicitation once if they wished to revise their threshold.

Post-vote survey. After the voting stage (and, for Threshold participants, the threshold-setting stage), all participants completed a brief post-vote survey. The survey collected the perceived social acceptability of voting for or against the proposal (on a Likert scale), self-reported engagement with the issue, political ideology, and demographics (gender identity, sexual orientation, race/ethnicity, year in school, and major/field of study).

Results distribution and realized disclosure. Once data collection was over, we emailed each participant a results summary. The email contained (i) aggregate vote shares for and against the proposal and the abstention rate, and (ii) a spreadsheet listing participants and their votes according to treatment-specific rules:

- In the *Private* treatment, the spreadsheet listed names only; no votes were shown.
- In the *Public* treatment, the spreadsheet listed each participant's name and, if they voted, their vote; abstainers appeared with no vote indicated.
- In the *Threshold* treatment, the spreadsheet listed all names; for non-abstainers whose realized like-minded vote share met or exceeded their chosen threshold t_i, the corresponding vote was displayed; all others appeared with no vote indicated.

This final distribution of results operationalizes the public signals Ψ^{Pri}, Ψ^{Pub}, and Ψ^{Thr} from the model.

3.1 Implementation Details

Pre-registration. We pre-registered our design and analysis plan on the AEA RCT Registry (AEARCTR-16968) prior to data collection. The final sample size is smaller than pre-specified because XLab was unable to recruit as many participants as it originally projected.

Sample composition. Our sample compromises 298 UC Berkeley undergraduate students recruited in October–November 2025. The sample has a median age of 20 years, with 68.5% identifying as female and 33.2% as non-heterosexual. Additional demographics appear in Online Appendix Table C1.

Video instructions. We delivered instructions via short videos (approximately 2 minutes for main instructions, 2.5 minutes for threshold-specific instructions) rather than text blocks. Video delivery standardizes the presentation of complex information, ensures participants cannot skip ahead without exposure to key details, and reduces attrition from text fatigue. Videos used simple animations with voice-over narration explaining the study context, voting options, and treatment-specific disclosure rules. Participants experiencing technical difficulties (e.g., audio issues) could access equivalent text instructions. Video transcripts, screens, and the survey instrument appear in Online Appendix D.

Elicitation procedure. Rather than asking participants to type a percentage or use a slider, we elicited thresholds through binary choices in a binary-search-style procedure. The initial value of the threshold was randomly drawn, and subsequent questions followed a standard binary-search logic. This approach (i) requires only simple binary comparisons and (ii) identifies thresholds on a fine grid with few questions.

Data quality. We follow best practices for online experiments (Haaland, Roth, and Wohlfart 2023). An attention check at the beginning of the survey screened out inattentive respondents. All remaining participants completed comprehension checks following the instructional videos; incorrect answers triggered immediate feedback and unlimited retries, and participants could proceed only after demonstrating full understanding. Online Appendix Table C3 summarizes data-quality metrics. Median completion time was 6.3 minutes, and first-pass comprehension rates were 61% for the voting task and 79% for the threshold task among Threshold participants.

Attrition and balance. Attrition in our experiment was modest at 9% and did not differ significantly by treatment (p=0.281). Online Appendix Table C2 shows that treatment groups are balanced on demographics, political orientation, and field of study.

Additional Details. Online Appendix C provides additional details on pre-registration, sample composition, and data quality.

4 Results

4.1 Treatment Effects on Participation and Expression

Figure 2 presents our main results. We first show that public observability distorts participation and expression, and then show that threshold majority voting eliminates these distortions.

Public voting distorts participation and expression. Panel A shows that public majority voting substantially increases abstention rates. In the Private treatment, 21.7% of participants abstain; in the Public treatment, this rate increases by 70%, rising to 37.1% ($p = 0.007$).

Panel B examines expression of the socially controversial view, reporting the fraction of all participants (including abstainers) in their treatment who voted for the controversial option. In the Private treatment, 43.4% do so; in the Public treatment, this fraction falls to 22.9% ($p = 0.001$). Thus, the expression of the controversial view through voting is almost twice as high in the Private Treatment as in the Public Treatment.

Panel C displays vote shares for the socially uncontroversial option conditional on participation. Among non-abstainers, 44.6% support the socially uncontroversial option in the Private treatment, compared to 63.6% in the Public treatment. Relative to anonymous voting, public voting thus increases the vote share for the socially uncontroversial option by around 40% ($p = 0.010$). In our setting, this shift is large enough that anonymous and public majority voting would lead to the implementation of opposite policy outcomes.

The comparison between the Private and Public aligns with prior empirical work (Braghieri 2024; Ho and Huang 2024) and with the equilibria described in the theoretical framework: public observability reduces participation, suppresses support for the socially controversial option, and shifts the vote margin toward the socially uncontroversial option.

Threshold voting restores truthful expression and reduces abstentions. Threshold majority voting eliminates the distortions of public voting. In the Threshold treatment, the abstention rate is 17.2% (Panel A), statistically indistinguishable from the 21.7% in the

Private treatment ($p = 0.437$).

Expression of the socially controversial view follows a similar pattern (Panel B). In the Threshold treatment, 42.5% of participants vote for the socially controversial option, a percentage that is statistically indistinguishable from the 43.4% of votes for the controversial option in Private ($p = 0.904$).

Panel C conditions on participation. Among non-abstainers, the fraction voting for the socially uncontroversial option in the Threshold treatment (48.6%) is statistically indistinguishable from its 44.6% counterpart from the Private treatment ($p = 0.618$).

Overall, in line with the theoretical model, threshold majority voting and anonymous majority voting cannot be statistically distinguished on all voting margins, thus corroborating hypotheses H1, H2, and H3. Threshold-Public comparisons also align with the model's predictions, with differences significant at the 1% level for abstention and expression of the controversial view, and at the 5% level for the uncontroversial vote share among non-abstainers.

4.2 Information Revelation and Strategic Disclosure

Threshold majority voting restores vote margins to the levels observed under anonymous majority voting. But does it also elicit the public expression of controversial views, or do participants simply set thresholds so high that no individual votes are disclosed? We find a meaningful degree of voluntary disclosure among both supporters and opponents, with threshold choices closely tracking the model's predictions.

Revelation patterns. Figure 3 presents revelation rates under the Threshold mechanism. In total, 33.3% of Threshold participants had their votes revealed—compared to 0% under Private and 100% of non-abstaining participants under Public (both by design).

Revelation is asymmetric by vote direction (Panel B). Among those voting for the socially uncontroversial option, 51.4% had their vote revealed; among those voting for the controversial option, only 29.7% did—a difference of 21.7 percentage points ($p = 0.031$). The fact that votes for both options are revealed confirms H4. The asymmetry reflects strategic threshold-setting, which we examine next.

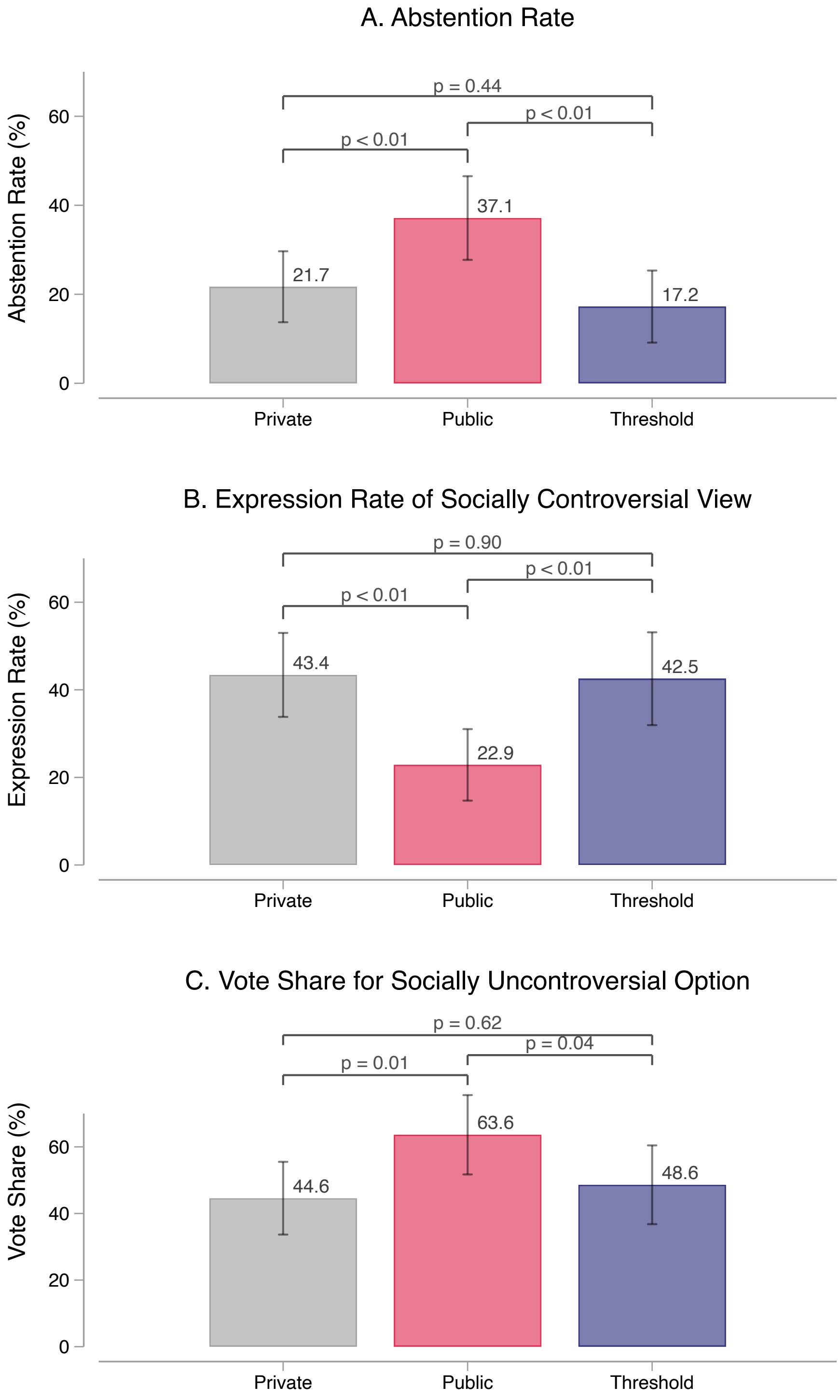


Figure 2: Treatment Effects on Voting Behavior

Note: This figure displays three voting outcomes by treatment. *Panel A: Abstention Rate.* The fraction of participants who chose to abstain rather than vote. *Panel B: Expression Rate of Socially Controversial View.* The fraction voting for the socially controversial option among all participants assigned to a treatment. *Panel C: Vote Share for Socially Uncontroversial Option.* Among participants who did not abstain, the fraction voting for the socially uncontroversial option in each treatment. Point estimates are sample means; error bars are 95% confidence intervals. Panels correspond to pre-registered hypotheses H1–H3, which predict Private = Threshold ≠ Public: Public increases abstention (H1) and uncontroversial vote share (H3), but decreases expression of the controversial view (H2); see Section 2. p-values from t-tests with robust standard errors: one-sided for directional hypotheses (Public vs. Private, Public vs. Threshold), two-sided for equality (Threshold vs. Private). N = 298.

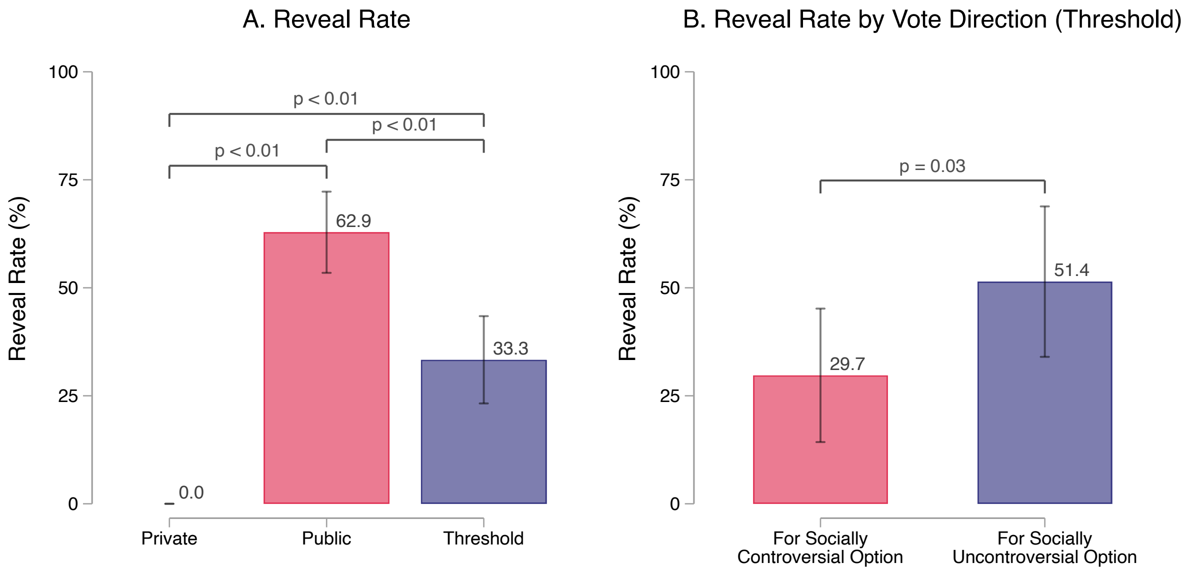


Figure 3: Vote Revelation Rates

Note: This figure displays revelation outcomes by treatment. *Panel A: Reveal Rate.* The fraction of participants whose votes were publicly revealed. In Private, no votes are revealed (by design). In Public, all non-abstaining votes are revealed (by design). In Threshold, a vote is revealed if the fraction voting the same way weakly exceeds the participant's chosen threshold. *Panel B: Reveal Rate by Vote Direction (Threshold).* Among Threshold non-abstainers, the fraction whose votes were revealed, by vote direction. That both rates are positive confirms H4; the asymmetry reflects strategic threshold-setting (H5). Point estimates are sample means; error bars are 95% confidence intervals. p-value from t-test with robust standard errors (one-sided, testing that uncontroversial reveal rate exceeds controversial).

Strategic threshold setting. Figure 4 plots the cumulative distribution of chosen thresholds as a function of the option supported by subjects in the Threshold treatment. The two distributions differ sharply. The median threshold is 45% among supporters of the socially uncontroversial option, compared to 75% among those supporting the socially controversial option. In the right tail, 75.7% of those voting for the socially controversial option set thresholds above 50% (requiring majority support before disclosure), compared to only 45.7% of those voting for the socially uncontroversial option.

These patterns are consistent with the model's prediction (H5): participants expressing the socially controversial view require larger fractions of like-minded voters before disclosing. We reject the null of equal distributions in favor of first-order stochastic dominance ($p = 0.011$; Goldman and Kaplan 2018).

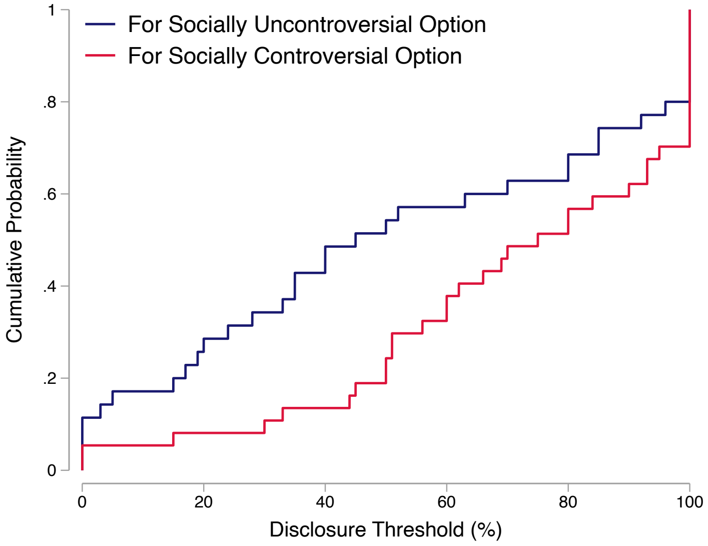


Figure 4: Strategic Threshold Setting by Vote Direction

Note: This figure shows the cumulative distribution function (CDF) of disclosure thresholds chosen by participants in the Threshold treatment. The x-axis is the threshold value: the minimum percentage of other non-abstaining participants who must vote the same way for the participant's vote to be publicly revealed. Two CDFs are plotted by vote direction: thresholds set by participants who voted for the socially uncontroversial option vs. for the socially controversial option. The rightward shift indicates that those voting for the controversial option set higher thresholds, consistent with hypothesis H5 (see Section 2). Median thresholds: 45% (uncontroversial) vs. 75% (controversial). Stochastic dominance test (Goldman and Kaplan 2018): $p = 0.011$.

4.3 Takeaways

Taken together, our results show that public observability in a politically sensitive campus setting generates the distortions highlighted in our motivating discussion and formal model. Moving from anonymous to public voting substantially raises abstention, suppresses expression of the socially controversial view, and shifts the vote margin toward the socially uncontroversial option, even though randomization fixes the underlying preference distribution. This mirrors the extensive- and intensive-margin distortions documented in prior work and captured by the public voting equilibrium in Proposition 1.

Threshold majority voting addresses these distortions while disclosing some information about individual behavior. On the core voting margins (participation, expression of the controversial view, and relative vote shares), the Threshold treatment is statistically indistinguishable from the Private treatment, corroborating hypotheses H1–H3 and the equilibrium characterization in Proposition 2. At the same time, a non-trivial share of both supporters and opponents of the proposal reveal their votes, and threshold choices are systematically higher among those holding the socially controversial position, consistent with H4–H5.

Substantively, the results suggest that threshold majority voting can reconcile two goals that often come into conflict in politically charged environments: encouraging truthful participation and preserving some scope for vote traceability.

4.4 Robustness

Comprehension, data quality, and inference. Online Appendix Table A1 reports treatment effects excluding potentially inattentive respondents: those who failed comprehension checks, the fastest 10% of respondents, and excluding the 6.9% of Threshold participants who revised their threshold choice. Treatment effects are stable across all specifications.

Alternative inference methods (multiple hypothesis correction, wild bootstrap, permutation tests; Online Appendix Table A2) yield consistent conclusions.

Heterogeneity. Online Appendix Figure A4 examines how public observability effects vary across subgroups (ideology, party, gender, issue engagement, LGBTQ+ status). Point estimates are broadly consistent, though women and highly engaged participants appear less susceptible to public observability distortions.

5 Conclusion

Many decision-making bodies face a tension between truthful expression, essential for aggregating preferences and information, and vote traceability, essential for assigning responsibility for collective decisions. When one policy option is socially stigmatized, fully anonymous procedures protect expression but obscure who backed which outcome, while fully public procedures provide a clean record of responsibility at the cost of distorting participation and choices. In this paper, we proposed and tested a simple mechanism—threshold majority voting—that addresses this tension by allowing individuals to decide under what aggregate conditions their vote becomes public.

After comparing the theoretical properties of anonymous majority voting, public majority voting, and threshold majority voting, we implemented the three mechanisms in an experiment at UC Berkeley on a contentious policy question: whether transgender women should be allowed to compete in women's collegiate sports. Moving from anonymous to public voting substantially increases abstention and nearly halves expression for the socially controversial option, shifting the vote margin toward the uncontroversial one. Threshold majority voting removes these distortions—abstention and vote shares closely mirror the anonymous benchmark—while revealing the votes of a meaningful share of participants. The mechanism therefore achieves the two desiderata of truthful voting and partial disclosure.

We acknowledge that both our theoretical and empirical analyses have limitations. The model is stylized, and the experiment is confined to a single issue, institution, and set of stakes. Future work could test threshold-based disclosure in other settings and compare it, theoretically or experimentally, to alternative forms of partial transparency.

Despite these limitations, our findings suggest that institutions need not choose starkly between secret and public ballots on sensitive issues. Threshold majority voting is a minor modification of majority rule that preserves the information content of anonymous voting while delivering meaningful vote traceability through selective disclosure. The mechanism may be particularly appealing to governance bodies, professional associations, and corporate boards that have to aggregate views on sensitive topics while navigating strong informal pressures. More broadly, the paper illustrates how formal institutional design can contain, rather than ignore, the influence of informal pressures on collective decisions.

References

Acemoglu, Daron, and Matthew O. Jackson. 2017. "Social Norms and the Enforcement of Laws." *Journal of the European Economic Association* 15 (2): 245–295.

Akbarpour, Mohammad, and Shengwu Li. 2020. "Credible Auctions: A Trilemma." *Econometrica* 88 (2): 425–467.

Ali, S. Nageeb, and Roland Bénabou. 2020. "Image versus Information: Changing Societal Norms and Optimal Privacy." *American Economic Journal: Microeconomics* 12 (3): 116–164.

Austen-Smith, David, and Jeffrey S. Banks. 1996. "Information Aggregation, Rationality, and the Condorcet Jury Theorem." *American Political Science Review* 90 (1): 34–45.

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

———. 2025. "Laws and Norms." *Journal of Political Economy,* Forthcoming.

Braghieri, Luca. 2024. "Political Correctness, Social Image, and Information Transmission." *American Economic Review* 114 (12): 3877–3904.

Brennan, Geoffrey, and Lauren Lomasky. 1993. *Democracy and Decision: The Pure Theory of Electoral Preference.* Cambridge: Cambridge University Press.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2020. "From Extreme to Mainstream: The Erosion of Social Norms." *American Economic Review* 110 (11): 3522–3548.

Bursztyn, Leonardo, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth. 2023. "Justifying Dissent." *The Quarterly Journal of Economics* 138 (3): 1403–1451.

Bursztyn, Leonardo, Georgy Egorov, and Robert Jensen. 2019. "Cool to Be Smart or Smart to Be Cool? Understanding Peer Pressure in Education." *The Review of Economic Studies* 86 (4): 1487–1526.

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 Robert Jensen. 2015. "How Does Peer Pressure Affect Educational Investments?" *The Quarterly Journal of Economics* 130 (3): 1329–1367.

———. 2017. "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure." *Annual Review of Economics* 9:131–153.

Condorcet, Nicolas de. 1785. *Essai Sur l'application de l'analyse à La Probabilité Des Décisions Rendues à La Pluralité Des Voix.* Paris: Imprimerie Royale.

Dellavigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao. 2017. "Voting to Tell Others." *The Review of Economic Studies* 84 (1): 143–181.

Downs, Anthony. 1957. "An Economic Theory of Political Action in a Democracy." *Journal of Political Economy* 65 (2): 135–150.

Feddersen, Timothy, and Wolfgang Pesendorfer. 1996. "The Swing Voter's Curse." *American Economic Review* 86 (3): 408–424.

———. 1997. "Voting Behavior and Information Aggregation in Elections With Private Information." *Econometrica* 65 (5): 1029–1058.

Funk, Patricia. 2010. "Social Incentives and Voter Turnout: Evidence from the Swiss Mail Ballot System." *Journal of the European Economic Association* 8 (5): 1077–1103.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102 (1): 33–48.

Gerber, Alan S., Gregory A. Huber, David Doherty, Conor M. Dowling, and Seth J. Hill. 2013. "Do Perceptions of Ballot Secrecy Influence Turnout? Results from a Field Experiment." *American Journal of Political Science* 57 (3): 537–551.

Goldman, Matt, and David M. Kaplan. 2018. "Comparing Distributions by Multiple Testing across Quantiles or CDF Values." *Journal of Econometrics* 206 (1): 143–166.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing Information Provision Experiments." *Journal of Economic Literature* 61 (1): 3–40.

Helmke, Gretchen, and Steven Levitsky. 2004. "Informal Institutions and Comparative Politics: A Research Agenda." *Perspectives on Politics* 2 (4): 725–740.

Ho, Yuen, and Yihong Huang. 2024. "Breaking the Spiral of Silence." *Working Paper.*

Kuran, Timur. 1987. "Chameleon Voters and Public Choice." *Public Choice* 53 (1): 53–78.

———. 1997. *Private Truths, Public Lies: The Social Consequences of Preference Falsification.* Cambridge, MA: Harvard University Press.

Levy, Gilat. 2007. "Decision Making in Committees: Transparency, Reputation, and Voting Rules." *American Economic Review* 97 (1): 150–168.

Maskin, Eric, and Jean Tirole. 2004. "The Politician and the Judge: Accountability in Government." *American Economic Review* 94 (4): 1034–1054.

Mattozzi, Andrea, and Marcos Y. Nakaguma. 2023. "Public versus Secret Voting in Committees." *Journal of the European Economic Association* 21 (3): 907–940.

Movement Advancement Project. 2026. *Bans on Transgender Youth Participation in Sports.*

Perez-Truglia, Ricardo, and Guillermo Cruces. 2017. "Partisan Interactions: Evidence from a Field Experiment in the United States." *Journal of Political Economy* 125 (4): 1208–1243.

Prat, Andrea. 2005. "The Wrong Kind of Transparency." *American Economic Review* 95 (3): 862–877.

Stevens, Sean T. 2025. *2026 College Free Speech Rankings: What Is the State of Free Speech on America's College Campuses?* Technical report. The Foundation for Individual Rights and Expression.

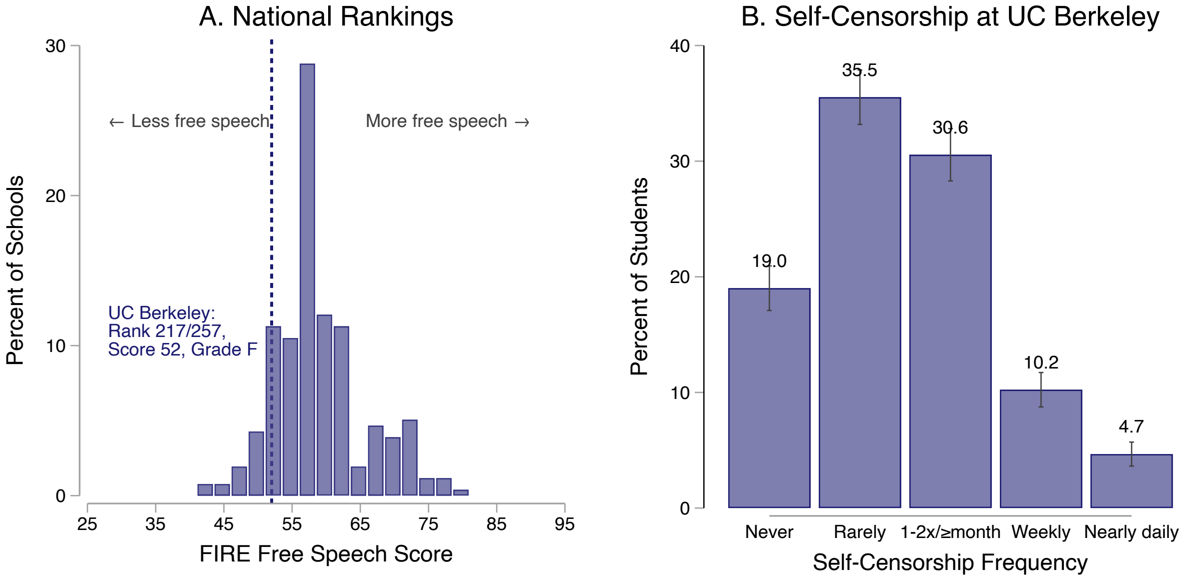
Visser, Bas, and Otto H. Swank. 2007. "On Committees of Experts." *The Quarterly Journal of Economics* 122 (1): 337–372.

Online Appendix:
Not for publication

Our supplementary material is structured as follows.

Section A contains additional figures and tables referenced in the text. Section B provides mathematical proofs for our motivating framework. Section C shows survey details and sample characteristics. Section D provides the survey instructions.

A Additional Figures and Tables



Appendix Figure A1: UC Berkeley's Position in National Free Speech Rankings

Note: This figure displays UC Berkeley's position in national free speech rankings and the prevalence of self-censorship among Berkeley students. *Panel A: National rankings.* The panel shows the distribution of overall free speech scores across 257 institutions in the 2026 FIRE College Free Speech Rankings (Stevens 2025). UC Berkeley's position (rank 217, score 52, grade F) is marked by the vertical line. The FIRE score aggregates institutional policies, administrative support for free expression, tolerance for controversial speakers, and student comfort expressing views. *Panel B: Self-censorship at UC Berkeley.* The bars display self-reported censorship frequency among Berkeley students from the 2026 FIRE campus speech survey (Stevens 2025), showing responses to: "How often, if ever, have you felt that you could not express your opinion on a subject because of how students, a professor, or the administration would respond?" (N = 1593, survey-weighted). Response categories range from "Never" to "Very often, nearly every day." Error bars are 95% confidence intervals.

Appendix Table A1: Treatment Effects under Alternative Sample Restrictions

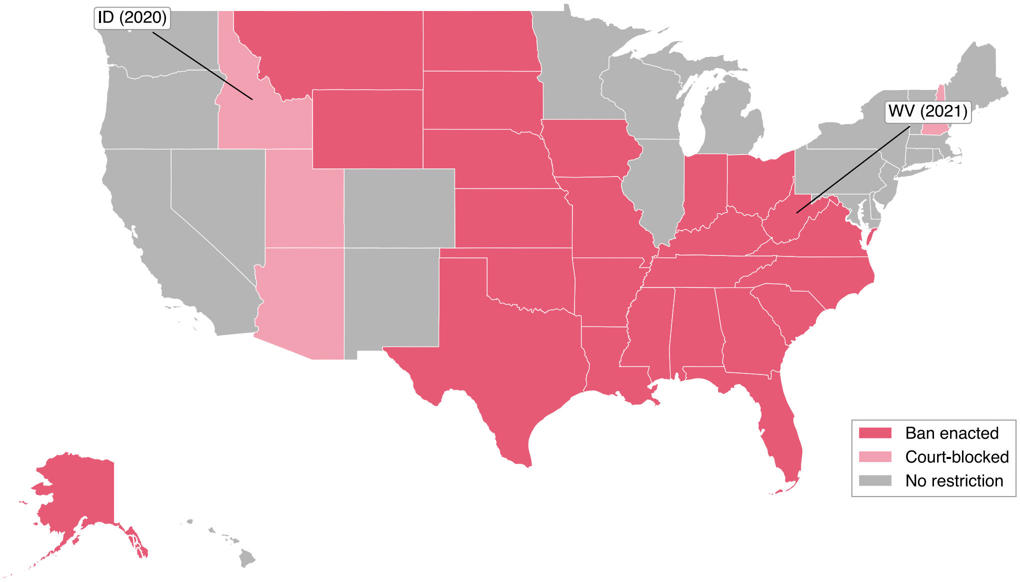
		Comprehension		Data Quality	
	Baseline	≤ 2 attempts	1 attempt	Excl. speeders	Excl. redo
	(1)	(2)	(3)	(4)	(5)
Panel A: Abstention Rate					
Public vs Private	15.44	14.73	14.27	15.34	15.44
	(6.22)	(6.43)	(7.95)	(6.72)	(6.22)
Threshold vs Private	−4.46	−3.01	−2.69	−4.50	−3.18
	(5.72)	(6.09)	(7.36)	(5.94)	(5.92)
Threshold vs Public	−19.90	−17.75	−16.97	−19.84	−18.62
	(6.25)	(6.61)	(8.12)	(6.57)	(6.43)
R^2	0.029	0.026	0.025	0.028	0.029
Observations	298	277 (93%)	182 (61%)	268 (90%)	292 (98%)
Panel B: Expression Rate of Socially Controversial View					
Public vs Private	−20.54	−20.55	−24.58	−20.97	−20.54
	(6.35)	(6.53)	(7.97)	(6.90)	(6.35)
Threshold vs Private	−0.87	−1.23	−3.20	−2.04	−2.66
	(7.20)	(7.48)	(9.13)	(7.45)	(7.32)
Threshold vs Public	19.67	19.32	21.39	18.93	17.88
	(6.73)	(7.01)	(8.43)	(6.99)	(6.86)
R^2	0.048	0.048	0.069	0.049	0.048
Observations	298	277 (93%)	182 (61%)	268 (90%)	292 (98%)
Panel C: Vote Share for Socially Uncontroversial Option					
Public vs Private	19.06	19.64	25.77	19.44	19.06
	(8.11)	(8.34)	(10.11)	(8.78)	(8.11)
Threshold vs Private	4.03	3.57	5.77	5.56	5.42
	(8.08)	(8.47)	(10.32)	(8.35)	(8.28)
Threshold vs Public	−15.03	−16.08	−20.00	−13.89	−13.64
	(8.41)	(8.78)	(10.58)	(8.82)	(8.61)
R^2	0.036	0.038	0.066	0.037	0.036
Observations	221	204 (92%)	136 (62%)	200 (90%)	215 (97%)

Note: This table reports treatment effects (percentage point differences) under alternative sample restrictions. Column (1): Baseline (main results from Figure 2). Columns (2)–(3): Comprehension restrictions (participants who passed comprehension checks within two attempts or on the first attempt). Columns (4)–(5): Data quality restrictions (excludes fastest 10% of respondents or Threshold participants who revised their threshold choice). Sample sizes shown at bottom of each panel. Panels correspond to hypotheses H1–H3, which predict Private = Threshold ≠ Public: Public increases abstention (H1) and uncontroversial vote share (H3), but decreases expression of the controversial view (H2). Robust standard errors in parentheses.

Appendix Table A2: Treatment Effects with Alternative Inference Methods

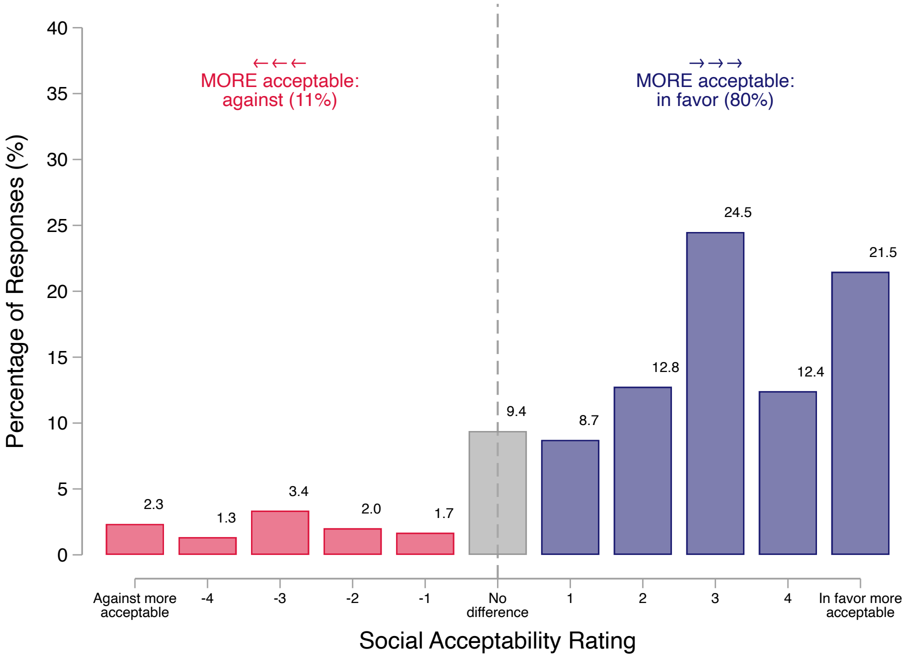
	Threshold vs Private (two-sided) (1)	Public vs Private (one-sided) (2)	Public vs Threshold (one-sided) (3)
Panel A: Abstention Rate			
Treatment effect	−4.46	15.44	−19.90
	(5.72)	(6.22)	(6.25)
Inference robustness:			
p-value: Robust	0.437	0.007	0.001
p-value: Romano-Wolf	0.714	0.016	0.005
p-value: Wild Bootstrap	0.426	0.008	0.000
p-value: Permutation	0.488	0.009	0.000
R^2	0.003	0.029	0.049
Observations	193	211	192
Panel B: Expression Rate of Socially Controversial View			
Treatment effect	−0.87	−20.54	19.67
	(7.20)	(6.35)	(6.73)
Inference robustness:			
p-value: Robust	0.904	0.001	0.002
p-value: Romano-Wolf	0.916	0.002	0.005
p-value: Wild Bootstrap	0.906	0.000	0.003
p-value: Permutation	1.000	0.002	0.003
R^2	0.000	0.048	0.044
Observations	193	211	192
Panel C: Vote Share for Socially Uncontroversial Option			
Treatment effect	4.03	19.06	−15.03
	(8.08)	(8.11)	(8.41)
Inference robustness:			
p-value: Robust	0.618	0.010	0.038
p-value: Romano-Wolf	0.733	0.016	0.035
p-value: Wild Bootstrap	0.607	0.010	0.037
p-value: Permutation	0.607	0.015	0.037
R^2	0.002	0.036	0.023
Observations	155	149	138

Note: This table reports treatment effects (percentage point differences) using four inference methods. Column (1): Threshold vs. Private (two-sided, testing equality). Column (2): Public vs. Private (one-sided, testing directional hypothesis). Column (3): Public vs. Threshold (one-sided, testing directional hypothesis). *Robust*: OLS with heteroskedasticity-robust standard errors (Figure 2). *Romano-Wolf*: familywise error rate control across three outcomes within each contrast, 1,000 bootstrap resamples. *Wild bootstrap*: heteroskedasticity-robust inference with Rademacher weights, 1,000 replications. *Permutation*: exact finite-sample randomization inference, 1,000 replications. Panel A tests H1 (abstention), Panel B tests H2 (expression), Panel C tests H3 (vote share).



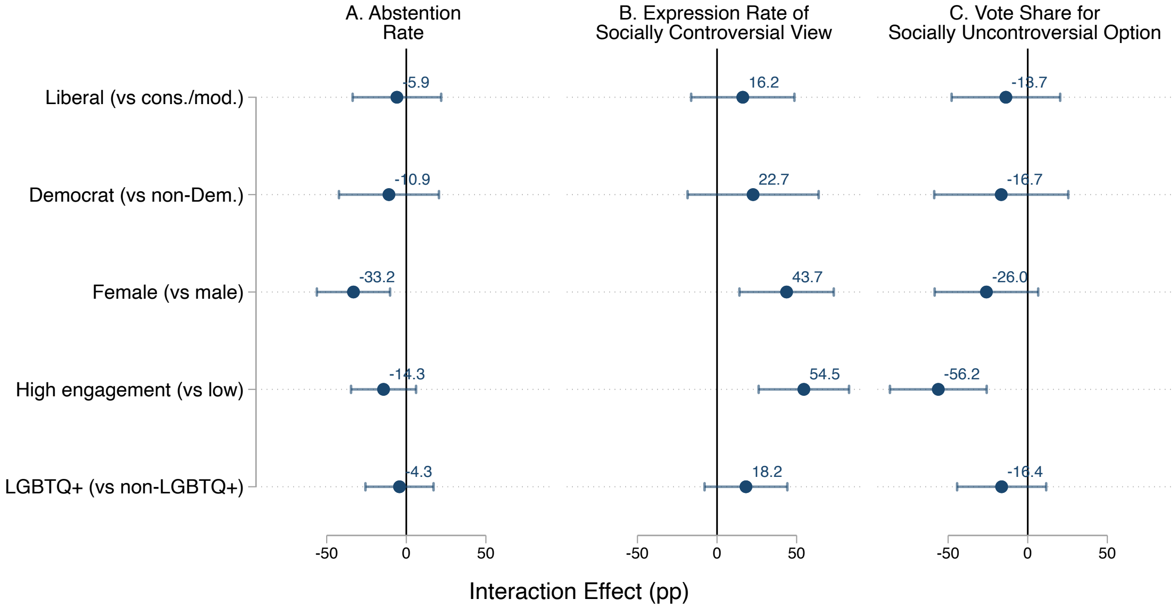
Appendix Figure A2: State Restrictions on Transgender Athlete Participation

Note: This figure displays U.S. states with laws or regulations restricting transgender athlete participation in school sports. Shaded states have enacted restrictions (29 states as of January 2026). Lighter shading indicates states where enforcement was blocked by court order (4 states: Idaho, Arizona, Utah, New Hampshire). Idaho (2020), the first state to enact such legislation, and West Virginia (2021) are annotated. Data from Movement Advancement Project (2026).



Appendix Figure A3: Perceived Social Acceptability of Each Position

Note: This figure displays participants' perceptions of the social acceptability of publicly expressing each position on the policy proposal (allowing transgender women to compete in women's collegiate sports). After voting, participants rated: "On this campus, do you think it's more socially acceptable to publicly say you're in favor of or against this proposal?" on an 11-point scale from -5 (much more acceptable to say you're *against*) to $+5$ (much more acceptable to say you're *in favor*), with 0 indicating no difference. The dashed vertical line marks the neutral point (0).



Appendix Figure A4: Public Observability Effects by Subgroup

Note: This figure displays public observability effects (Public − Private) across participant subgroups. *Panel A: Abstention.* Percentage-point difference in abstention. Positive values indicate public observability increases abstention. *Panel B: Expression.* Percentage-point difference in expression of the socially controversial view (voting against transgender women in collegiate sports). Negative values indicate public observability suppresses expression. *Panel C: Vote share.* Percentage-point difference in the socially uncontroversial vote share among non-abstainers. Positive values indicate public observability shifts votes toward the uncontroversial option. Each coefficient is β_3 from $Y_i = \beta_0 + \beta_1 \text{Public}_i + \beta_2 \text{Subgroup}_i + \beta_3(\text{Public}_i \times \text{Subgroup}_i) + \varepsilon_i$, estimated with robust standard errors. Horizontal bars are 95% confidence intervals; the vertical line at zero indicates no differential effect. Subgroup definitions and sample sizes: Liberal (N=225) vs. conservative/moderate (N=73); Democrat (N=268) vs. non-Democrat (N=30); Female (N=204) vs. male (N=81; other/non-binary excluded); High engagement (N=130) vs. low (N=168); LGBTQ+ (N=99) vs. non-LGBTQ+ (N=199).

B Mathematical Appendix

B.1 Proof of Proposition 1

We prove Proposition 1 by means of two lemmas, each characterizing the equilibrium under one of the two voting mechanisms.

Because the population is a continuum, no individual agent is pivotal. Consequently, the instrumental term $\beta\,\mathbb{E}_i[\mathbf{1}\{\bar{a}=\omega\}]$ does not affect any agent's marginal incentives and can be treated as a constant. Without loss of generality, we therefore set $\beta = 0$ in what follows.

Lemma 1 (Anonymous Majority Voting)**.** *There exists $\bar{c} > 0$ such that, if $c_H > \bar{c}$, the following holds. Under anonymous majority voting there exists a unique equilibrium in which all low-cost types vote truthfully and all high-cost types abstain.*

Proof. Under anonymous majority voting, the public signal Ψ^{Pri} contains only aggregate outcomes (abstention rate and vote shares) and never identifies any individual's action. Because the population is a continuum, a single agent has measure zero and her action does not affect these aggregates. Hence her action does not affect the audience's posteriors about a "named" agent: for any type $\boldsymbol{x}_i$ and any $a_i \in \{0, 1, \tilde{a}\}$,

$$\mathbb{E}_i\big[P_j(a_i=\omega \mid \Psi^{Pri}, \omega)\big] \quad \text{and} \quad \mathbb{E}_i\big[f(\mu(\Psi))\,P_j(a_i=1 \mid \Psi^{Pri}, \omega)\big]$$

are constant in a_i. Moreover, under anonymous voting the privacy term is always zero, since individual votes are never revealed. The only components of u_i that depend on a_i are therefore the expressive benefit $\phi\mathbf{1}\{a_i = s_i\}$ and the participation cost $c_i\mathbf{1}\{a_i \neq \tilde{a}\}$.

Consider first a low-cost type with $c_i = 0$. For such an agent,

$$u_i(\boldsymbol{x}_i, a_i = s_i) = (\text{constant}) + \phi, \qquad u_i(\boldsymbol{x}_i, a_i \neq s_i) = (\text{constant}), \qquad u_i(\boldsymbol{x}_i, \tilde{a}) = (\text{constant}).$$

Since $\phi > 0$, voting truthfully strictly dominates both misvoting and abstaining, so every low-cost type uniquely best responds by choosing $a_i = s_i$.

Next consider a high-cost type with $c_i = c_H$. If she abstains, she pays no participation cost and receives no expressive benefit, so her payoff is some constant K. If she votes truthfully, her payoff is $K + \phi - c_H$; if she misvotes, it is at most $K - c_H$. We can pick $c_H > \phi$, so $\phi - c_H < 0$, and thus any action $a_i \in \{0, 1\}$ yields strictly lower utility than abstaining. Hence every high-cost type strictly prefers $a_i = \tilde{a}$.

Thus, in any equilibrium all low-cost types vote truthfully and all high-cost types abstain. This profile is the unique equilibrium. □

Lemma 2 (Public Majority Voting)**.** *There exist $\bar{c}, \bar{\pi}, \bar{\phi}, \underline{\eta} > 0$ such that, if $c_H > \bar{c}$, $\pi_H > \bar{\pi}$, $\phi > \bar{\phi}$, and $\eta_H > \underline{\eta}$, the following holds. There exists an equilibrium under public majority voting in which agents' strategies are*

$$a_i^{Pub} = \begin{cases} \tilde{a} & \textit{if } (c_i = c_H) \ \vee \ (c_i = 0 \wedge \pi_i = \pi_H), \\ 1 & \textit{if } (c_i = 0 \wedge \pi_i = 0 \wedge \eta_i = 0 \wedge s_i = 1), \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. Under public majority voting, the public signal Ψ^{Pub} reveals, for each agent, whether she abstained and, if she voted, which option she chose. Thus

$$\mathbf{1}\{\Psi^{Pub} \text{ reveals } a_i\} = \mathbf{1}\{a_i \in \{0, 1\}\},$$

and abstainers never pay the privacy cost. Because the population is a continuum, a single agent has measure zero and her action does not affect aggregate vote shares or the behavior of others. Her action affects only her own label in Ψ^{Pub} (”abstained”, ”voted 0”, ”voted 1”), and thus only her own stigma, privacy, and reputation-for-accuracy terms.

Under public voting, an abstainer has no expressive benefit, no stigma or privacy cost, and is not seen as having supported either policy; thus her payoff is

$$u_i(\boldsymbol{x}_i, \tilde{a}) = 0 \qquad \text{for all } \boldsymbol{x}_i.$$

We verify this is an equilibrium by checking best responses type-by-type.

1. High-cost types $(c_i = c_H)$**.**

Fix any $(s_i, \eta_i, \pi_i, c_H)$. If i abstains, her payoff is $u_i(\boldsymbol{x}_i, \tilde{a}) = 0$. If she votes, her most favorable case is: she votes truthfully ($a_i = s_i$), suffers no stigma ($\eta_i = 0$) and no privacy cost ($\pi_i = 0$), and enjoys the maximal reputation-for-accuracy gain $\kappa \Pr(\omega = s_i \mid s_i) = \kappa\lambda$. In this best case,

$$u_i(\boldsymbol{x}_i, s_i) \leq \phi + \kappa\lambda - c_H.$$

We can pick $c_H > \phi + \kappa\lambda$, so $\phi + \kappa\lambda - c_H < 0$. Any other vote choice yields weakly lower payoff (because it removes expressive benefits and/or adds stigma and privacy costs). Hence, for all high-cost types,

$$u_i(\boldsymbol{x}_i, \tilde{a}) > \max\{u_i(\boldsymbol{x}_i, 0), u_i(\boldsymbol{x}_i, 1)\},$$

so $a_i^{Pub} = \tilde{a}$ is a strict best response for all $c_i = c_H$.

2. Low-cost, high-privacy types ($c_i = 0,\ \pi_i = \pi_H$)**.**

Fix any $(s_i, \eta_i, \pi_H, 0)$. If i abstains, $u_i(\boldsymbol{x}_i, \tilde{a}) = 0$. If she votes truthfully, her best-case payoff (with zero stigma) is

$$u_i(\boldsymbol{x}_i, s_i) \leq \phi + \kappa\lambda - \pi_H.$$

We can pick $\pi_H > \phi + \kappa\lambda$, so this upper bound is strictly negative. Any other vote (e.g. $a_i \neq s_i$) yields at most the reputation-for-accuracy term $\kappa(1 - \lambda)$ plus the same privacy cost, and is therefore even less attractive. Hence, for all low-cost, high-privacy types, voting is strictly dominated by abstaining, and $a_i^{Pub} = \tilde{a}$ is a strict best response.

3. Low-cost, low-privacy, low-stigma types ($c_i = 0,\ \pi_i = 0,\ \eta_i = 0$)**.**

For these types the stigma and privacy terms vanish. Their payoff from voting depends only on expressive benefits and reputation-for-accuracy.

Case $s_i = 0$.

If i votes $a_i = 0$, she obtains

$$\begin{aligned} u_i((0,0,0,0),0) &= \phi\,\mathbf{1}\{a_i = s_i\} + \kappa \Pr(\omega = 0 \mid s_i = 0) \\ &= \phi + \kappa\lambda. \end{aligned}$$

If she votes $a_i = 1$, she obtains

$$u_i((0,0,0,0),1) = 0 + \kappa \Pr(\omega = 1 \mid s_i = 0) = \kappa(1-\lambda).$$

If she abstains, $u_i((0,0,0,0),\tilde{a}) = 0$.

Since $\lambda > 1/2$ and $\phi > 0$,

$$u_i((0,0,0,0),0) - u_i((0,0,0,0),1) = \phi + \kappa(2\lambda - 1) > 0,$$

and

$$u_i((0,0,0,0),0) - u_i((0,0,0,0),\tilde{a}) = \phi + \kappa\lambda > 0.$$

Thus when $(s_i, \eta_i, \pi_i, c_i) = (0,0,0,0)$ a strict best response is $a_i^{Pub} = 0$.

Case $s_i = 1$.

If i votes $a_i = 1$, she obtains

$$u_i((1,0,0,0),1) = \phi + \kappa \Pr(\omega = 1 \mid s_i = 1) = \phi + \kappa\lambda.$$

If she votes $a_i = 0$, she obtains

$$u_i((1,0,0,0),0) = 0 + \kappa \Pr(\omega = 0 \mid s_i = 1) = \kappa(1-\lambda),$$

and if she abstains, $u_i((1,0,0,0),\tilde{a}) = 0$.

Again,

$$u_i((1,0,0,0),1) - u_i((1,0,0,0),0) = \phi + \kappa(2\lambda - 1) > 0,$$

and

$$u_i((1,0,0,0),1) - u_i((1,0,0,0),\tilde{a}) = \phi + \kappa\lambda > 0.$$

Hence when $(s_i, \eta_i, \pi_i, c_i) = (1,0,0,0)$ a strict best response is $a_i^{Pub} = 1$.

Thus all low-cost, low-privacy, low-stigma types vote truthfully: $a_i^{Pub} = s_i$.

4. Low-cost, low-privacy, high-stigma types ($c_i = 0,\ \pi_i = 0,\ \eta_i = \eta_H$)**.**

For η_H sufficiently large, high-stigma types vote $a_i^{Pub} = 0$ regardless of s_i.

Case $s_i = 0$.

If i votes $a_i = 0$, she gets the same payoff as the corresponding low-stigma type (since stigma only applies to $a_i = 1$):

$$u_i((0, \eta_H, 0, 0), 0) = \phi + \kappa\lambda.$$

If she votes $a_i = 1$, she gains no expressive benefit, earns reputation-for-accuracy $\kappa(1 - \lambda)$, and suffers a non-negative stigma cost. Even ignoring stigma, her payoff from $a_i = 1$ is at most

$$u_i((0, \eta_H, 0, 0), 1) \leq \kappa(1 - \lambda).$$

Abstaining yields 0. Therefore

$$u_i((0, \eta_H, 0, 0), 0) > \max\{u_i((0, \eta_H, 0, 0), 1),\ u_i((0, \eta_H, 0, 0), \tilde{a})\},$$

so when $s_i = 0$ a strict best response is $a_i^{Pub} = 0$.

Case $s_i = 1$.

For type $(1, \eta_H, 0, 0)$ under the candidate profile:

If she votes $a_i = 0$, she obtains no expressive benefit, no stigma, and reputation-for-accuracy payoff

$$u_i((1, \eta_H, 0, 0), 0) = \kappa \Pr(\omega = 0 \mid s_i = 1) = \kappa(1 - \lambda).$$

As before, this strictly dominates abstention:

$$u_i((1, \eta_H, 0, 0), 0) - u_i((1, \eta_H, 0, 0), \tilde{a}) = \kappa(1 - \lambda) > 0.$$

If instead she votes $a_i = 1$, her individual vote is publicly observed and the audience infers $a_i = 1$ with probability one, so $P_j(a_i = 1 \mid \Psi^{Pub}, \omega) = 1$. Her expected stigma cost is therefore

$$\eta_H \, \mathbb{E}_i\Big[f(\mu(\Psi)) \ \Big| \ s_i = 1, a_i = 1\Big].$$

Under the candidate strategy profile, the set of non-abstainers coincides with $\{c_i = 0, \pi_i = 0\}$, and among these non-abstainers:

- Low-stigma types ($\eta_i = 0$) vote $a_i = s_i$.
- High-stigma types ($\eta_i = \eta_H$) vote $a_i = 0$ regardless of s_i.

Let $p_{\eta_L} := \Pr(\eta_i = \eta_L)$ denote the population share of low-stigma types. By independence of (η_i, π_i, c_i) and s_i, and since all $c_i = 0, \pi_i = 0$ types vote under the candidate profile, the equilibrium share of votes for $a = 1$ among non-abstainers in state ω, denoted $\mu_1(\omega)$, is

$$\begin{aligned}
\mu_1(\omega) &= \Pr(a_i = 1 \mid c_i = 0, \pi_i = 0, \omega) \\
&= \Pr(\eta_i = \eta_L \mid c_i = 0, \pi_i = 0) \Pr(s_i = 1 \mid \omega) \\
&= p_{\eta_L} \Pr(s_i = 1 \mid \omega)
\end{aligned}$$

so that

$$\mu_1(0) = p_{\eta_L}(1 - \lambda), \qquad \mu_1(1) = p_{\eta_L}\lambda.$$

In a continuum, aggregate vote shares are deterministic conditional on the state, so $\mu(\Psi^{Pub}) = \mu_1(\omega)$ almost surely given ω. Conditional on $s_i = 1$, the posterior probability of each state is

$$\Pr(\omega = 0 \mid s_i = 1) = 1 - \lambda, \qquad \Pr(\omega = 1 \mid s_i = 1) = \lambda.$$

Therefore

$$\begin{aligned}
\mathbb{E}_i\Big[f(\mu(\Psi)) \;\Big|\; s_i = 1, a_i = 1\Big] &= \sum_{\omega \in \{0,1\}} \Pr(\omega \mid s_i = 1) f(\mu(\Psi)) \\
&= (1 - \lambda) \min\Big\{M, \frac{1}{p_{\eta_L}(1 - \lambda)}\Big\} + \lambda \min\Big\{M, \frac{1}{p_{\eta_L}\lambda}\Big\}.
\end{aligned}$$

Since $\lambda > 1/2$, we have $\lambda > 1 - \lambda$ and therefore

$$\frac{1}{p_{\eta_L}\lambda} < \frac{1}{p_{\eta_L}(1 - \lambda)}.$$

Hence

$$S^{Pub} := \begin{cases} M, & \text{if } M \leq \dfrac{1}{p_{\eta_L}\lambda}, \\ (1-\lambda)M + \dfrac{1}{p_{\eta_L}}, & \text{if } \dfrac{1}{p_{\eta_L}\lambda} < M \leq \dfrac{1}{p_{\eta_L}(1-\lambda)}, \\ \dfrac{2}{p_{\eta_L}}, & \text{if } M > \dfrac{1}{p_{\eta_L}(1-\lambda)}. \end{cases}$$

In all cases $S^{Pub} > 0$. The expected stigma cost from choosing $a_i = 1$ is then $\eta_H S^{Pub}$, and the corresponding payoff from $a_i = 1$ satisfies

$$u_i((1, \eta_H, 0, 0), 1) = \phi + \kappa\lambda - \eta_H S^{Pub}.$$

Comparing $a_i = 1$ and $a_i = 0$, we obtain

$$\begin{aligned} u_i((1, \eta_H, 0, 0), 1) - u_i((1, \eta_H, 0, 0), 0) &= \phi + \kappa\lambda - \eta_H S^{Pub} - \kappa(1-\lambda) \\ &= \phi + \kappa(2\lambda - 1) - \eta_H S^{Pub}. \end{aligned}$$

Define

$$\underline{\eta}^{Pub} := \frac{\phi + \kappa(2\lambda - 1)}{S^{Pub}}.$$

Since $\lambda \in (1/2, 1)$ and $\phi > 0$, the numerator is strictly positive, so $\underline{\eta}^{Pub} > 0$. If $\eta_H > \underline{\eta}^{Pub}$, then

$$\phi + \kappa(2\lambda - 1) - \eta_H S^{Pub} < 0,$$

which implies

$$u_i((1, \eta_H, 0, 0), 1) < u_i((1, \eta_H, 0, 0), 0).$$

Combining this with $u_i((1, \eta_H, 0, 0), 0) > u_i((1, \eta_H, 0, 0), \tilde{a})$, we conclude that for all $\eta_H > \underline{\eta}^{Pub}$ a strict best response of this type is $a_i^{Pub} = 0$.

5. Equilibrium existence.

Steps 1–4 show that the profile is an equilibrium. □

Together, Lemmas 1 and 2 prove Proposition 1.

□

B.2 Proof of Proposition 2

We first establish two general properties of the threshold majority voting mechanism (Lemmas 3 and 4) that hold for any equilibrium and simplify the subsequent analysis. We then characterize the unique equilibrium (Lemma 5). The rest of the proof pins down the parameter regions where both Proposition 1 and Proposition 2 hold simultaneously.

Lemma 3 (Canonical Thresholds)**.** *Fix an arbitrary equilibrium of the threshold majority voting mechanism. For each option $a \in \{0, 1\}$ and state $\omega \in \{0, 1\}$, let $\mu_a(\omega) \in [0, 1]$ denote the equilibrium share of non-abstainers who vote for option a in state ω. Define*

$$\mu_a^{\min} := \min\{\mu_a(0), \mu_a(1)\}, \qquad \mu_a^{\max} := \max\{\mu_a(0), \mu_a(1)\}.$$

Consider any agent i who, given her type, decides to vote for option a and chooses a threshold $t_i \in [0, 1]$. Then:

(i) For any two thresholds $t, t' \in [0, 1]$ such that $\mathbf{1}\{\mu_a(0) \geq t\} = \mathbf{1}\{\mu_a(0) \geq t'\}$ and $\mathbf{1}\{\mu_a(1) \geq t\} = \mathbf{1}\{\mu_a(1) \geq t'\}$, the agent's expected utility satisfies $u_i(\boldsymbol{x}_i, a, t) = u_i(\boldsymbol{x}_i, a, t')$.

(ii) For $t < 1$, the revelation pattern $R_a(t) := \big(\mathbf{1}\{\mu_a(0) \geq t\}, \mathbf{1}\{\mu_a(1) \geq t\}\big)$ is constant on each of three regions: $R_a(t) = (1, 1)$ for $t \in [0, \mu_a^{\min}]$; $R_a(t) \in \{(0, 1), (1, 0)\}$ for $t \in (\mu_a^{\min}, \mu_a^{\max}]$; and $R_a(t) = (0, 0)$ for $t \in (\mu_a^{\max}, 1]$.

(iii) By the tie-breaking rule (agents choose the largest threshold when indifferent), any optimal threshold for side a can be represented by one of the three canonical thresholds*: $t_a^{\text{low}} := \mu_a^{\min}$, $t_a^{\text{int}} := \mu_a^{\max}$, and $t_a^{\text{high}} := 1$.*

Proof. Intuitively, an agent's payoff depends on her threshold choice only through *when* her vote is revealed—in neither state, one state, or both. This observation reduces the infinite threshold space to just three payoff-relevant choices.

Fix an equilibrium and an agent i who chooses to vote for option a and to set a threshold $t_i \in [0, 1]$. Recall the disclosure rule: for $t_i < 1$, i's individual vote is revealed in state ω if and only if $\mu_a(\omega) \geq t_i$. For $t_i = 1$, the mechanism gives a "never reveal" option.

Part (i): Utility depends only on the revelation pattern. Define $R_a(t) := \big(\mathbf{1}\{\mu_a(0) \geq t\}, \mathbf{1}\{\mu_a(1) \geq t\}\big)$ for $t < 1$, and $R_a(1) = (0, 0)$. We show that $u_i(\boldsymbol{x}_i, a, t)$ depends on t only through $R_a(t)$. The expressive and participation terms depend on a_i but not t_i. The privacy term depends on whether i is revealed in each state, which is encoded by $R_a(t_i)$. The stigma and reputation-for-accuracy terms depend on the audience's beliefs about i, which—since i has measure zero and cannot affect aggregates—depend only on whether i's vote is revealed in each state. Thus if $R_a(t) = R_a(t')$, then $u_i(\boldsymbol{x}_i, a, t) = u_i(\boldsymbol{x}_i, a, t')$.

Part (ii): Three regions. By definition of $\mu_a^{\min}$ and $\mu_a^{\max}$: if $t \leq \mu_a^{\min}$, both inequalities $\mu_a(\omega) \geq t$ hold, so $R_a(t) = (1, 1)$; if $t \in (\mu_a^{\min}, \mu_a^{\max}]$, exactly one holds; if $t > \mu_a^{\max}$, neither holds, so $R_a(t) = (0, 0)$.

Part (iii): Canonical thresholds. By parts (i) and (ii), utility is constant within each region. By the tie-breaking convention, the agent chooses the largest threshold in her preferred region: $t_a^{\text{low}} := \mu_a^{\min}$, $t_a^{\text{int}} := \mu_a^{\max}$, or $t_a^{\text{high}} := 1$. □

Lemma 4 (Label Structure)**.** *Fix an arbitrary equilibrium of the threshold majority voting mechanism. For each agent i, the public signal Ψ^{Thr} induces a* personal label *$L_i \in \{L^0, L^1, L^u\}$: L^0 if i is revealed as voting $a_i = 0$, L^1 if revealed as voting $a_i = 1$, and L^u if undisclosed. Then:*

(i) For payoff purposes, the audience's beliefs about i depend on Ψ^{Thr} only through (L_i, ω). Specifically, the reputation-for-accuracy and stigma terms can be written as functions of the label-conditional posteriors $P_j(a_i = \cdot \mid L, \omega)$ and $S_L(\omega)$.

(ii) For revealing labels: $P_j(a_i = 0 \mid L^0, \omega) = 1$, $P_j(a_i = 1 \mid L^1, \omega) = 1$, $S_{L^0}(\omega) = 0$, and $S_{L^1}(\omega) \in [0, M]$ for all ω.

(iii) Any deviation by i that preserves her label in both states leaves the label-conditional beliefs $P_j(a_i = \cdot \mid L, \omega)$ and $S_L(\omega)$ unchanged.

Proof. The mechanism partitions agents into three observationally distinct groups: those revealed as voting $a = 0$, those revealed as voting $a = 1$, and those who remain undisclosed. These labels exhaust all individual-level information available to the audience.

Part (i): By iterated expectations, conditioning on L_i suffices.

Part (ii): Immediate from the definitions.

Part (iii): Since i has measure zero, her deviation cannot affect aggregates; if the label is unchanged, so are label-conditional expectations. □

Lemma 5 (Threshold Majority Voting)**.** *There exist $\bar{p}_{c_H}, \bar{c}, \bar{\pi}, \bar{\phi}, \underline{\eta}, \bar{\eta} > 0$ with $\underline{\eta} < \bar{\eta}$ such that, if $p_{c_H} > \bar{p}_{c_H}$, $c_H > \bar{c}$, $\pi_H > \bar{\pi}$, $\phi > \bar{\phi}$, and $\eta_H \in (\underline{\eta}, \bar{\eta})$, the following holds. Under threshold majority voting there exists a unique equilibrium in which agents' strategies are*

$$(a_i^{Thr}, t_i) = \begin{cases} \tilde{a} & \textit{if } c_i = c_H, \\ (s_i, 1) & \textit{if } c_i = 0 \wedge \pi_i = \pi_H, \\ (s_i, \lambda) & \textit{if } c_i = 0 \wedge \pi_i = 0 \wedge \eta_i = 0, \\ (0, 1-\lambda) & \textit{if } c_i = 0 \wedge \pi_i = 0 \wedge \eta_i = \eta_H \wedge s_i = 0, \\ (1, 1) & \textit{if } c_i = 0 \wedge \pi_i = 0 \wedge \eta_i = \eta_H \wedge s_i = 1. \end{cases}$$

Proof. Because the population is a continuum, no individual agent is pivotal. Consequently, the instrumental term $\beta\,\mathbb{E}_i[\mathbf{1}\{\bar{a} = \omega\}]$ does not affect any agent's marginal incentives and can be treated as a constant. Without loss of generality, we therefore set $\beta = 0$ in what follows.

The proof proceeds as follows. By Lemma 3, an agent's payoff depends on her threshold choice only through a coarse revelation pattern, reducing the infinite threshold space to three canonical choices. By Lemma 4, the audience's beliefs about each agent depend only on a simple three-valued label. Given these simplifications, we characterize voting behavior (Sublemma B.1) and threshold choices (Sublemma B.2) type by type. Finally, we verify that the parameter restrictions are mutually compatible.

Sub-lemma B.1. *There exist $\bar{c}, \bar{\pi}, \bar{\phi}, \underline{\eta}, \bar{\eta} > 0$ with $\underline{\eta} < \bar{\eta}$ such that, if $c_H > \bar{c}$, $\pi_H > \bar{\pi}$, $\phi > \bar{\phi}$, and $\eta_H \in (\underline{\eta}, \bar{\eta})$, the following holds. In all equilibria under threshold majority voting, agents with $c_i = c_H$ abstain and all other agents vote truthfully.*

Proof. We verify voting behavior type by type, beginning with high-cost types who abstain regardless of other characteristics (Step 1.A), then proceeding through increasingly complex cases: high-privacy types (Step 1.B), low-stigma types (Step 1.C), and finally high-stigma types by signal (Steps 1.D and 1.E).

Step 1.A: High-cost types abstain.

Claim. For any type with $c_i = c_H$ and any (a_i, t_i) with $a_i \in \{0,1\}$: $u_i(\boldsymbol{x}_i, a_i, t_i) < u_i(\boldsymbol{x}_i, \tilde{a})$. Abstention is strictly dominant for high-cost types.

Proof. Fix a type $\boldsymbol{x}_i = (s_i, \eta_i, \pi_i, c_H)$ and some voting strategy (a_i, t_i) with $a_i \in \{0,1\}$ and $t_i \in [0,1]$. Define the payoff difference between voting and abstaining as

$$\Delta u_i(a_i, t_i) := u_i(\boldsymbol{x}_i, a_i, t_i) - u_i(\boldsymbol{x}_i, \tilde{a}).$$

We bound $\Delta u_i(a_i, t_i)$ from above term by term.

Expressive benefit. When abstaining, $a_i = \tilde{a}$ and the expressive term is

$$\phi\, \mathbf{1}\{\tilde{a} = s_i\} = 0.$$

When voting, the expressive term is $\phi\, \mathbf{1}\{a_i = s_i\}$, which is either 0 or ϕ. Hence the best-case expressive gain from voting is

$$\phi\, \mathbf{1}\{a_i = s_i\} - 0 \;\le\; \phi.$$

reputation-for-accuracy. The reputation-for-accuracy term is

$$\kappa\, \mathbb{E}_i\big[P_j(a_i = \omega \mid \Psi^{Thr}, \omega)\big],$$

where the posterior $P_j(a_i = \omega \mid \Psi^{Thr}, \omega)$ always lies in $[0,1]$. Changing from abstaining to voting therefore changes this expectation by at most 1 in absolute value. Thus the reputation-for-accuracy gain (i.e. the increase in utility due to this term) is bounded by

$$\Delta(\text{reputation-for-accuracy}) \;\le\; \kappa.$$

Stigma. For any action (voting or abstaining) we have

$$-\eta_i\, \mathbb{E}_i\Big[f(\mu(\Psi))\, P_j(a_i = 1 \mid \Psi^{Thr}, \omega)\Big].$$

The factor $f(\mu(\Psi))$ is in $[0, M]$ and $P_j(a_i = 1 \mid \Psi^{Thr}, \omega) \in [0, 1]$, so the product inside the expectation lies in $[0, M]$. Hence, for any strategy,

$$-M \leq -\eta_i \, \mathbb{E}_i\Big[f(\mu(\Psi)) \, P_j(a_i = 1 \mid \Psi^{Thr}, \omega)\Big] \leq 0.$$

In particular, for any two actions (including abstention and any vote), the change in the stigma term is bounded in absolute value by $\eta_i M$:

$$\Big| -\eta_i \, \mathbb{E}_i\Big[f(\mu(\Psi)) \, P_j(a_i = 1 \mid \Psi^{Thr}, \omega)\Big] \Big| \ \leq \ \eta_i M.$$

Hence the *most* that i can gain from stigma by switching from abstention to voting is at most $\eta_i M \leq \eta_H M$:

$$\Delta(\text{stigma}) \ \leq \ \eta_H M.$$

Privacy. The privacy term is

$$-\pi_i \, \mathbb{E}_i\big[\mathbf{1}\{a_i \in \{0, 1\} \text{ and } \Psi^{Thr} \text{ reveals } a_i\}\big].$$

When abstaining, we have $a_i = \tilde{a}$ and the indicator is zero, so the privacy term is exactly 0. When voting, the indicator $\mathbf{1}\{\Psi^{Thr}$ reveals $a_i\}$ is weakly positive, and the term is weakly *more* negative than under abstention. Thus privacy contributes a weakly non-positive amount to $\Delta u_i(a_i, t_i)$:

$$\Delta(\text{privacy}) \ \leq \ 0.$$

Participation cost. The participation cost is $-c_i \, \mathbf{1}\{a_i \neq \tilde{a}\}$. When abstaining, this cost is 0. When voting, since $a_i \in \{0, 1\}$, the cost is $-c_H$. Therefore participation contributes exactly

$$\Delta(\text{participation}) = -c_H$$

to $\Delta u_i(a_i, t_i)$.

Combining the bounds. Summing the contributions of all terms, we obtain the upper bound

$$\Delta u_i(a_i,t_i) \;\le\; \underbrace{\phi}_{\text{expressive}} + \underbrace{\kappa}_{\text{reputation-for-accuracy}} + \underbrace{\eta_H M}_{\text{stigma}} + \underbrace{0}_{\text{privacy}} + \underbrace{(-c_H)}_{\text{participation}} = \phi + \kappa + \eta_H M - c_H.$$

We can pick $c_H > \phi + \kappa + \eta_H M$, so

$$\phi + \kappa + \eta_H M - c_H \;<\; 0.$$

Hence $\Delta u_i(a_i,t_i) < 0$ for every choice of (a_i,t_i) with $a_i \in \{0,1\}$ and $t_i \in [0,1]$.

Abstention is strictly dominant for high-cost types. □

Step 1.B: High-privacy types never reveal and vote truthfully.

Claim. For types with $c_i = 0$, $\pi_i = \pi_H$, in any equilibrium:

(i) For every action $a_i \in \{0,1\}$, the unique optimal (canonical) threshold on side a_i is $t_i = 1$ (the "never reveal" option).

(ii) Given $t_i = 1$, truthful voting $a_i = s_i$ strictly dominates both misreporting and abstaining. Hence the unique best response of such types is $(a_i^{Thr}, t_i) = (s_i, 1)$.

Proof.

(i) Optimal threshold choice: $t_i = 1$. Fix $a_i \in \{0,1\}$. By Lemma 3, we restrict attention to the three canonical thresholds:

$$t_{a_i}^{\text{low}} = \mu_{a_i}^{\min}, \qquad t_{a_i}^{\text{int}} = \mu_{a_i}^{\max}, \qquad t_{a_i}^{\text{high}} = 1,$$

corresponding to reveal in both states, reveal in one state, and never reveal, respectively.

Low threshold $t_{a_i}^{\text{low}}$. If $t_i = t_{a_i}^{\text{low}} = \mu_{a_i}^{\min}$, then by the definition of $\mu_{a_i}^{\min}$ the inequality $\mu_{a_i}(\omega) \ge t_i$ holds in both states $\omega = 0, 1$. Thus i is revealed whenever she votes a_i, so

$$\mathbb{E}_i\big[\mathbf{1}\{\Psi^{Thr} \text{ reveals } a_i\}\big] = 1.$$

The expected privacy cost is therefore

$$-\pi_H \, \mathbb{E}_i\big[\mathbf{1}\{\Psi^{Thr} \text{ reveals } a_i\}\big] = -\pi_H.$$

Intermediate threshold $t_{a_i}^{\text{int}}$. If $t_i = t_{a_i}^{\text{int}} = \mu_{a_i}^{\max}$, then i is revealed exactly in the state ω^* where $\mu_{a_i}(\omega^*) = \mu_{a_i}^{\max}$, and is undisclosed in the other state. Conditional on her private signal s_i, the posterior probability that $\omega = \omega^*$ is either λ or $1-\lambda$, depending on which state has higher support for a_i. Hence

$$\mathbb{E}_i\big[\mathbf{1}\{\Psi^{Thr} \text{ reveals } a_i\}\big] = \Pr_i(\omega = \omega^* \mid s_i) \in \{\lambda, 1-\lambda\} \;\geq\; 1-\lambda,$$

because $\lambda > 1/2$. Thus the expected privacy cost under any intermediate threshold satisfies

$$-\pi_H \, \mathbb{E}_i\big[\mathbf{1}\{\Psi^{Thr} \text{ reveals } a_i\}\big] \;\leq\; -\pi_H(1-\lambda).$$

High threshold $t_{a_i}^{\text{high}} = 1$. If $t_i = 1$, the mechanism never reveals i's vote, so

$$\mathbb{E}_i\big[\mathbf{1}\{\Psi^{Thr} \text{ reveals } a_i\}\big] = 0,$$

and the expected privacy cost is exactly 0.

Comparison. For fixed a_i, changing t_i affects only privacy, stigma, and reputation-for-accuracy (since $c_i = 0$).

The reputation-for-accuracy term has the form

$$\kappa \, \mathbb{E}_i\big[P_j(a_i = \omega \mid \Psi^{Thr}, \omega)\big],$$

with $P_j(a_i = \omega \mid \Psi^{Thr}, \omega) \in [0,1]$, so for any two thresholds t, t' the change in this term is at most κ in absolute value.

The stigma term change between any two thresholds is at most $\eta_H M$ in absolute value. Thus switching to $t_i = 1$ can decrease the combined (stigma+reputation-for-accuracy) payoff by at most $\kappa + \eta_H M$.

Moving from any canonical $t_i < 1$ to $t_i = 1$ increases the privacy payoff by at least $\pi_H(1-\lambda)$. Therefore,

$$u_i(\boldsymbol{x}_i, a_i, 1) - u_i(\boldsymbol{x}_i, a_i, t_i) \ \geq \ \pi_H(1-\lambda) - (\kappa + \eta_H M).$$

Picking $\pi_H > (\kappa+\eta_H M)/(1-\lambda)$ ensures $u_i(\boldsymbol{x}_i, a_i, 1) > u_i(\boldsymbol{x}_i, a_i, t_i)$ for every canonical $t_i < 1$. Thus the unique optimal threshold is $t_i = 1$.

(ii) Action choice given $t_i = 1$: truthful voting. Under $t_i = 1$, i's label is L^u for any action $a_i \in \{0, 1, \tilde{a}\}$. By Lemma 4, the reputation-for-accuracy and stigma terms are identical for all actions yielding L^u.

Since $c_i = 0$ and privacy cost is zero when $t_i = 1$, the only term depending on a_i is the expressive benefit:

$$u_i(\boldsymbol{x}_i, a_i, 1) = \phi\, \mathbf{1}\{a_i = s_i\} + \kappa\, \mathbb{E}_i[P_j(a_i = \omega \mid L^u, \omega)] - \eta_i\, \mathbb{E}_i\Big[f(\mu(\Psi))P_j(a_i = 1 \mid L^u)\Big],$$

while abstaining yields

$$u_i(\boldsymbol{x}_i, \tilde{a}) = \kappa\, \mathbb{E}_i[P_j(a_i = \omega \mid L^u, \omega)] - \eta_i\, \mathbb{E}_i\Big[f(\mu(\Psi))P_j(a_i = 1 \mid L^u)\Big].$$

Taking differences, $u_i(\boldsymbol{x}_i, a_i, 1) - u_i(\boldsymbol{x}_i, \tilde{a}) = \phi\, \mathbf{1}\{a_i = s_i\}$. Hence truthful voting yields $\phi > 0$ relative to abstaining, while misreporting yields zero gain. Combining (i) and (ii), the unique best response is $(a_i^{Thr}, t_i) = (s_i, 1)$.

Step 1.C: Low-stigma types vote truthfully.

For a type $\boldsymbol{x}_i = (s_i, 0, 0, 0)$, the expected utility from choosing (a_i, t_i) under the threshold mechanism is

$$u_i((s_i, 0, 0, 0), a_i, t_i) = \underbrace{\phi\, \mathbf{1}\{a_i = s_i\}}_{\text{expressive}} + \underbrace{\kappa\, \mathbb{E}_i\big[P_j(a_i = \omega \mid \Psi^{Thr}, \omega)\big]}_{\text{reputation-for-accuracy}},$$

since privacy, stigma, and participation costs are all zero.

Conditioning on the state and using the fact that, given s_i, $\Pr(\omega = s_i \mid s_i) = \lambda$ and

$\Pr(\omega = 1 - s_i \mid s_i) = 1 - \lambda$, we can rewrite the reputation-for-accuracy term as

$$\kappa\,\mathbb{E}_i\big[P_j(a_i = \omega \mid \Psi^{Thr}, \omega)\big] = \kappa\lambda\,\mathbb{E}_i\big[P_j(a_i = s_i \mid \Psi^{Thr}, \omega = s_i)\big] + \kappa(1-\lambda)\,\mathbb{E}_i\big[P_j(a_i = 1 - s_i \mid \Psi^{Thr}, \omega = 1 - s_i)\big]$$

Hence

$$u_i((s_i,0,0,0), a_i, t_i) = \phi\,\mathbf{1}\{a_i = s_i\} + \kappa\lambda\,\mathbb{E}_i\big[P_j(a_i = s_i \mid \Psi^{Thr}, \omega = s_i)\big] \tag{2}$$

$$+\kappa(1-\lambda)\,\mathbb{E}_i\big[P_j(a_i = 1 - s_i \mid \Psi^{Thr}, \omega = 1 - s_i)\big]. \tag{3}$$

We now show that, for such types, truthful voting strictly dominates both misreporting and abstention. This step can be easily done by requiring $\bar{\phi} \geq \kappa$. However, we establish an even lower bound, namely $\bar{\phi} \geq k(1-\lambda)$ that will be useful later in the proof.

(i) Truthful voting vs. misreporting.

Fix a realization of the signal $s_i \in \{0,1\}$ for agent i and consider two pure strategies for this type:

$$\sigma^{\mathrm{tr}} := (a_i, t_i) = (s_i, t_i^{\mathrm{tr}}), \qquad \sigma^{\mathrm{mr}} := (a_i, t_i) = (1 - s_i, t_i^{\mathrm{mr}}),$$

where $t_i^{\mathrm{tr}}, t_i^{\mathrm{mr}} \in [0,1]$ are arbitrary thresholds. We compare the expected utilities $u_i((s_i,0,0,0), \sigma^{\mathrm{tr}})$ and $u_i((s_i,0,0,0), \sigma^{\mathrm{mr}})$.

From (2), the expressive terms differ by exactly ϕ:

$$\phi\,\mathbf{1}\{a_i = s_i\} = \begin{cases} \phi & \text{under } \sigma^{\mathrm{tr}}, \\ 0 & \text{under } \sigma^{\mathrm{mr}}. \end{cases}$$

Thus

$$u_i((s_i,0,0,0), \sigma^{\mathrm{tr}}) - u_i((s_i,0,0,0), \sigma^{\mathrm{mr}}) = \phi + \Delta\mathrm{Acc},$$

where $\Delta\mathrm{Acc}$ is the difference in reputation-for-accuracy payoffs:

$$\Delta\mathrm{Acc} := \kappa\lambda\,\Delta_s + \kappa(1-\lambda)\,\Delta_{1-s},$$

with

$$\Delta_s := \mathbb{E}_i\big[P_j(a_i = s_i \mid \Psi^{Thr}, \omega = s_i)\big]_{\sigma^{\text{tr}}} - \mathbb{E}_i\big[P_j(a_i = s_i \mid \Psi^{Thr}, \omega = s_i)\big]_{\sigma^{\text{mr}}},$$

$$\Delta_{1-s} := \mathbb{E}_i\big[P_j(a_i = 1 - s_i \mid \Psi^{Thr}, \omega = 1 - s_i)\big]_{\sigma^{\text{tr}}} - \mathbb{E}_i\big[P_j(a_i = 1 - s_i \mid \Psi^{Thr}, \omega = 1 - s_i)\big]_{\sigma^{\text{mr}}}.$$

The subscripts indicate which strategy (σ^{tr} or σ^{mr}) is used by agent i when computing the expectation.

We now bound Δ_s and Δ_{1-s} using only the label structure from Lemma 4 and the fact that posteriors are probabilities in $[0, 1]$.

State $\omega = s_i$.

In this state, under σ^{tr} the agent votes $a_i = \omega$, while under σ^{mr} she votes $a_i = 1 - \omega$. Let $P^{\text{tr}}(L \mid \omega = s_i)$ and $P^{\text{mr}}(L \mid \omega = s_i)$ denote the probabilities (under the two strategies) that the agent receives label $L \in \{L^0, L^1, L^u\}$ in state $\omega = s_i$.

By the definition of labels and Lemma 4:

$$P_j(a_i = s_i \mid L^{s_i}, \omega = s_i) = 1, \qquad P_j(a_i = s_i \mid L^{1-s_i}, \omega = s_i) = 0,$$

and there exists some $q_s \in [0, 1]$ such that

$$P_j(a_i = s_i \mid L^u, \omega = s_i) = q_s,$$

where q_s depends only on the equilibrium belief system and not on i's individual deviation.

Under σ^{tr}, the agent can receive labels L^{s_i} or L^u, but never L^{1-s_i}, so

$$\mathbb{E}_i\big[P_j(a_i = s_i \mid \Psi^{Thr}, \omega = s_i)\big]_{\sigma^{\text{tr}}} = P^{\text{tr}}(L^{s_i} \mid \omega = s_i) \cdot 1 + P^{\text{tr}}(L^u \mid \omega = s_i) \cdot q_s.$$

Under σ^{mr}, the agent can receive labels L^{1-s_i} or L^u, but never L^{s_i}, so

$$\mathbb{E}_i\big[P_j(a_i = s_i \mid \Psi^{Thr}, \omega = s_i)\big]_{\sigma^{\text{mr}}} = P^{\text{mr}}(L^{1-s_i} \mid \omega = s_i) \cdot 0 + P^{\text{mr}}(L^u \mid \omega = s_i) \cdot q_s.$$

Taking the difference, we obtain

$$\begin{aligned}\Delta_s &= \mathbb{E}_i\big[P_j(a_i = s_i \mid \Psi^{Thr}, \omega = s_i)\big]_{\sigma^{\text{tr}}} - \mathbb{E}_i\big[P_j(a_i = s_i \mid \Psi^{Thr}, \omega = s_i)\big]_{\sigma^{\text{mr}}} = \\ &= P^{\text{tr}}(L^{s_i} \mid \omega = s_i)\,(1 - q_s) + P^{\text{mr}}(L^{1-s_i} \mid \omega = s_i)\,q_s.\end{aligned}$$

Since $P^{\text{tr}}(\cdot \mid \omega = s_i)$ and $P^{\text{mr}}(\cdot \mid \omega = s_i)$ are probabilities and $q_s \in [0,1]$, both coefficients $1 - q_s$ and q_s are non-negative. Hence $\Delta_s \geq 0$.

State $\omega = 1 - s_i$.

In this state, under σ^{tr} the agent votes $a_i = s_i \neq \omega$, while under σ^{mr} she votes $a_i = \omega$. The random variables

$$P_j(a_i = 1 - s_i \mid \Psi^{Thr}, \omega = 1 - s_i)$$

are probabilities and therefore lie in $[0,1]$ under either strategy. It follows directly that their expectations under the two strategies differ by at most 1 in absolute value:

$$-1 \;\leq\; \Delta_{1-s} := \mathbb{E}_i\big[P_j(a_i = 1{-}s_i \mid \Psi^{Thr}, \omega = 1{-}s_i)\big]_{\sigma^{\text{tr}}} - \mathbb{E}_i\big[P_j(a_i = 1{-}s_i \mid \Psi^{Thr}, \omega = 1{-}s_i)\big]_{\sigma^{\text{mr}}} \;\leq\; 1.$$

In particular,

$$\Delta_{1-s} \;\geq\; -1.$$

Bounding the reputation-for-accuracy gain from misreporting.

Putting the two states together, we have

$$\Delta_s \geq 0, \qquad \Delta_{1-s} \geq -1.$$

Therefore

$$\Delta\text{Acc} = \kappa\lambda\,\Delta_s + \kappa(1-\lambda)\,\Delta_{1-s} \;\geq\; \kappa\lambda \cdot 0 + \kappa(1-\lambda)\cdot(-1) = -\kappa(1-\lambda),$$

or equivalently

$$\kappa\,\mathbb{E}_i\big[P_j(a_i = \omega \mid \Psi^{Thr}, \omega)\big]_{\sigma^{\text{mr}}} - \kappa\,\mathbb{E}_i\big[P_j(a_i = \omega \mid \Psi^{Thr}, \omega)\big]_{\sigma^{\text{tr}}} \;\leq\; \kappa(1-\lambda).$$

Thus the maximal reputation-for-accuracy *gain* from switching from truthful voting to misreporting is bounded above by $\kappa(1-\lambda)$.

Returning to the utility difference,

$$u_i((s_i,0,0,0),\sigma^{\mathrm{tr}}) - u_i((s_i,0,0,0),\sigma^{\mathrm{mr}}) = \phi + \Delta\mathrm{Acc} \ \geq \ \phi - \kappa(1-\lambda).$$

We can pick $\phi > \kappa(1-\lambda)$, so

$$u_i((s_i,0,0,0),\sigma^{\mathrm{tr}}) - u_i((s_i,0,0,0),\sigma^{\mathrm{mr}}) > 0$$

for every misreporting strategy σ^{mr} and every threshold choice t_i^{tr} under truthful voting. Hence, for low-stigma low-privacy low-cost types, truthful voting $a_i = s_i$ strictly dominates misreporting $a_i = 1 - s_i$ for any thresholds.

(ii) Truthful voting vs. abstention.

We now show that truthful voting strictly dominates abstention for at least one threshold, namely $t_i = 1$.

Fix $t_i = 1$ and consider two actions for type $(s_i, 0, 0, 0)$:

$$\sigma^{\mathrm{tr},1} := (a_i, t_i) = (s_i, 1), \qquad \sigma^{\mathrm{abs}} := (a_i, t_i) = (\tilde{a}, \cdot).$$

Under $\sigma^{\mathrm{tr},1}$ the agent votes but is never individually revealed, so her label is L^u in both states. Under σ^{abs}, she also appears as “undisclosed”, i.e. her label is again L^u in both states. By Lemma 4, the reputation-for-accuracy term depends only on (L_i, ω), so the reputation-for-accuracy payoff is *identical* under $\sigma^{\mathrm{tr},1}$ and σ^{abs}.

On the other hand, the expressive term is ϕ under $\sigma^{\mathrm{tr},1}$ and 0 under σ^{abs}. Thus

$$u_i((s_i,0,0,0),\sigma^{\mathrm{tr},1}) - u_i((s_i,0,0,0),\sigma^{\mathrm{abs}}) = \phi > 0.$$

Therefore, there exists a threshold (in fact, $t_i = 1$) such that truthful voting strictly dominates abstention for low-stigma low-privacy low-cost types.

(iii) Conclusion. Truthful voting strictly dominates both misreporting (for any thresholds)

and abstaining (with $t_i = 1$). Hence such types vote truthfully in every equilibrium. □

Combining Steps 1.B and 1.C: in every equilibrium, low-cost high-privacy types and low-cost low-privacy low-stigma types all vote truthfully. The mass of $a_i = 1$ votes from these groups in state ω is

$$\big(p_{c_L} p_{\pi_H} + p_{c_L} p_{\pi_L} p_{\eta_L}\big) \Pr(s_i = 1 \mid \omega).$$

Let

$$q^{\text{tr}} := p_{c_L} p_{\pi_H} + p_{c_L} p_{\pi_L} p_{\eta_L} > 0.$$

Hence, conditional on any state ω,

$$\mu(\Psi^{Thr} \mid \omega) \;\geq\; q^{\text{tr}} \Pr(s_i = 1 \mid \omega) \;\geq\; q^{\text{tr}}(1-\lambda).$$

Define

$$\mu^{Thr}_{\min} := q^{\text{tr}}(1-\lambda) = \big(p_{c_L} p_{\pi_H} + p_{c_L} p_{\pi_L} p_{\eta_L}\big)(1-\lambda).$$

Then in any equilibrium and for each $\omega \in \{0,1\}$ we have

$$\mu(\Psi^{Thr} \mid \omega) \;\geq\; \mu^{Thr}_{\min} > 0. \tag{4}$$

In particular,

$$0 < \mu^{Thr}_{\min} \leq \mu(\Psi^{Thr}) \leq 1, \quad \Rightarrow \quad \min\Big\{M, \frac{1}{\mu(\Psi^{Thr})}\Big\} \leq \min\Big\{M, \frac{1}{\mu^{Thr}_{\min}}\Big\}.$$

Note that $P_j(a_i = 1 \mid L^u, \omega) \in (0,1)$ for each ω, since both votes appear among undisclosed agents.

Step 1.D: High-stigma, low-privacy, low-cost types with $s_i = 0$ vote truthfully.

Proof. Consider type $(0, \eta_H, 0, 0)$. We rule out abstention, then misreporting.

(i) Abstention is strictly dominated by truthful voting.

Consider two strategies for this type:

$$\sigma^{\text{tr},1} := (a_i, t_i) = (0, 1), \qquad \sigma^{\text{abs}} := (a_i, t_i) = (\tilde{a}, \cdot).$$

Under $\sigma^{\text{tr},1}$ the agent never reveals her vote and therefore receives the undisclosed label $L_i = L^u$ in both states; under σ^{abs} she is also undisclosed and again receives label $L_i = L^u$ in both states. Because each agent has measure zero, the composition of the L^u pool and the aggregate vote shares are unaffected by i's unilateral deviation in a_i. By Lemma 4, the reputation-for-accuracy and stigma terms depend only on (L_i, ω), so they are identical under $\sigma^{\text{tr},1}$ and σ^{abs}. For this type we have $c_i = 0$ and $\pi_i = 0$, so participation and privacy costs are zero under both strategies as well.

Hence the only payoff difference between $\sigma^{\text{tr},1}$ and σ^{abs} is the expressive term:

$$u_i((0, \eta_H, 0, 0), \sigma^{\text{tr},1}) - u_i((0, \eta_H, 0, 0), \sigma^{\text{abs}}) = \phi\, \mathbf{1}\{a_i = s_i\}_{\sigma^{\text{tr},1}} - \phi\, \mathbf{1}\{a_i = s_i\}_{\sigma^{\text{abs}}} = \phi > 0.$$

Thus abstention is strictly dominated by truthful voting with $t_i = 1$, and this type must participate in any equilibrium.

(ii) Misreporting $a_i = 1$ is strictly dominated by truthful voting.

We now compare truthful voting with $a_i = 0$ to misreporting $a_i = 1$. Let $t_0^{\text{low}} := \mu_0^{\text{min}}$ denote the canonical low threshold on side 0 from Lemma 3. Consider the two strategies

$$\sigma^{\text{tr,low}} := (a_i, t_i) = (0, t_0^{\text{low}}), \qquad \sigma^{\text{mr}} := (a_i, t_i) = (1, t_i^{\text{mr}}),$$

where $t_i^{\text{mr}} \in [0, 1]$ is an arbitrary threshold.

Under $\sigma^{\text{tr,low}}$ the agent is revealed as having voted $a_i = 0$ in both states, so her label is $L_i = L^0$ and, by Lemma 4,

$$P_j(a_i = 1 \mid L^0, \omega) = 0 \quad \Rightarrow \quad \text{Stigma}_i(\sigma^{\text{tr,low}}) = 0.$$

Under σ^{mr} she votes $a_i = 1$, so her label can never be L^0: depending on t_i^{mr} and the state, she is either revealed as L^1 or remains undisclosed L^u. From the analysis above (using the presence of high-privacy truthful types and the bounds on $\mu(\Psi^{Thr})$), we know that the expected stigma term under σ^{mr} is strictly negative:

$$\text{Stigma}_i(\sigma^{\text{mr}}) = -\eta_H\, \mathbb{E}_i\Big[\min\{M, \frac{1}{\mu}(\Psi^{Thr})\}\, P_j(a_i = 1 \mid \Psi^{Thr}, \omega)\Big] < 0.$$

Therefore the stigma contribution to the difference $u_i(\sigma^{\text{tr,low}}) - u_i(\sigma^{\text{mr}})$ is strictly positive:

$$\Delta\text{Stigma} := \text{Stigma}_i(\sigma^{\text{tr,low}}) - \text{Stigma}_i(\sigma^{\text{mr}}) = 0 - \text{Stigma}_i(\sigma^{\text{mr}}) > 0.$$

Next, consider the expressive and reputation-for-accuracy terms. For the type $(s_i = 0, \eta_i = 0, \pi_i = 0, c_i = 0)$, Step 1.C established that for any pair of strategies

$$\sigma^{\text{tr}} = (a_i, t_i) = (s_i, t_i^{\text{tr}}), \qquad \sigma^{\text{mr}} = (a_i, t_i) = (1 - s_i, t_i^{\text{mr}}),$$

the reputation-for-accuracy gain from misreporting is bounded above by $\kappa(1-\lambda)$. The same bound on the reputation-for-accuracy difference applies here, since the reputation-for-accuracy term does not depend on η_i. Thus, for our $(0, \eta_H, 0, 0)$ type and the strategies $\sigma^{\text{tr,low}}$ and σ^{mr},

$$\left[u_i(\sigma^{\text{tr,low}}) - u_i(\sigma^{\text{mr}})\right]_{\text{expr+acc}} \geq \phi - \kappa(1-\lambda),$$

where the subscript indicates we restrict attention to the expressive plus reputation-for-accuracy components. We can pick $\phi > \kappa(1-\lambda)$, so this part of the difference is strictly positive.

Combining the expressive, reputation-for-accuracy, and stigma components, we obtain

$$u_i((0, \eta_H, 0, 0), \sigma^{\text{tr,low}}) - u_i((0, \eta_H, 0, 0), \sigma^{\text{mr}}) > \left[\phi - \kappa(1-\lambda)\right] + 0 \; > \; 0.$$

Since t_i^{mr} was arbitrary, misreporting $a_i = 1$ is strictly dominated by truthful voting $a_i = 0$ for this type.

(iii) Conclusion. Abstention and misreporting are both strictly dominated by truthful voting. Hence such types vote $a_i = 0$ in every equilibrium. □

Step 1.E: High-stigma, low-privacy, low-cost types with $s_i = 1$ vote truthfully.

This is the crux of the argument. High-stigma agents with signal $s_i = 1$ would, under public voting, be tempted to vote against their signal to avoid stigma. Threshold voting breaks this tension: by choosing a high disclosure threshold, they can vote truthfully while limiting stigma exposure.

Claim. There exist $\bar{c}, \bar{\pi}, \bar{\phi}, \underline{\eta}, \bar{\eta} > 0$ with $\underline{\eta} < \bar{\eta}$ such that, if $c_H > \bar{c}$, $\pi_H > \bar{\pi}$, $\phi > \bar{\phi}$, and $\eta_H \in (\underline{\eta}, \bar{\eta})$, the following holds. In any equilibrium of the threshold majority voting mechanism, any agent of type

$$\boldsymbol{x}_i = (s_i, \eta_i, \pi_i, c_i) = (1, \eta_H, 0, 0)$$

(i.e. low participation cost, low privacy cost, high stigma, signal $s_i = 1$) must participate and vote truthfully, $a_i = 1$.

Proof. Fix an arbitrary equilibrium of the threshold mechanism and a type $\boldsymbol{x}_i = (1, \eta_H, 0, 0)$.

<u>(a) Participation: ruling out abstention.</u>

Consider the two strategies

$$\sigma^{\text{tr},1} := (a_i, t_i) = (1, 1), \qquad \sigma^{\text{abs}} := (a_i, t_i) = (\tilde{a}, \cdot).$$

Under $\sigma^{\text{tr},1}$ the agent votes for $a_i = 1$ but never reveals her individual vote and therefore receives the undisclosed label $L_i = L^u$ in every realization of the public signal. Under σ^{abs} she abstains and is also undisclosed, so her label is again $L_i = L^u$ in every realization. Because the agent has measure zero, her deviation does not affect aggregate vote shares or the distribution of the public signal Ψ^{Thr}.

By Lemma 4, the reputation-for-accuracy and stigma terms depend only on (L_i, ω) (and the aggregate behavior), not on the agent's own individual action. Therefore, under $\sigma^{\text{tr},1}$ and σ^{abs} the reputation-for-accuracy, stigma, privacy, and participation terms coincide. Since $c_i = \pi_i = 0$, the only payoff difference between these two strategies arises from the expressive term:

$$u_i((1, \eta_H, 0, 0), \sigma^{\text{tr},1}) - u_i((1, \eta_H, 0, 0), \sigma^{\text{abs}}) = \phi > 0.$$

Hence abstention is strictly dominated by truthful voting with $t_i = 1$ for this type, independently of η_H. In particular, any best response for this type must involve participation.

<u>(b) Vote choice: ruling out $a_i = 0$.</u>

Now compare a truthful strategy with $a_i = 1$ to misreporting strategies with $a_i = 0$. Let

$\sigma^{\text{tr}} = (a_i, t_i) = (1, t_i^{\text{tr}})$ be an arbitrary truthful strategy (with an arbitrary threshold), and let $\sigma^{\text{mr}} = (a_i, t_i) = (0, t_i^{\text{mr}})$ be an arbitrary misreporting strategy. For each such pair, we decompose the utility difference as

$$u_i((1, \eta_H, 0, 0), \sigma^{\text{tr}}) - u_i((1, \eta_H, 0, 0), \sigma^{\text{mr}}) = \underbrace{\big[\phi + \Delta\text{Acc}\big]}_{\text{expressive + reputation-for-accuracy}} - \eta_H \, \Delta S,$$

where

$$\Delta\text{Acc} := \kappa \, \mathbb{E}_i\big[P_j(a_i = \omega \mid \Psi^{Thr}, \omega)\big]_{\sigma^{\text{tr}}} - \kappa \, \mathbb{E}_i\big[P_j(a_i = \omega \mid \Psi^{Thr}, \omega)\big]_{\sigma^{\text{mr}}},$$

$$\Delta S := \mathbb{E}_i\Big[f(\mu(\Psi))P_j(a_i = 1 \mid \Psi^{Thr}, \omega)\Big]_{\sigma^{\text{tr}}} - \mathbb{E}_i\Big[f(\mu(\Psi))P_j(a_i = 1 \mid \Psi^{Thr}, \omega)\Big]_{\sigma^{\text{mr}}}.$$

reputation-for-accuracy term. By the argument in Step 1.C (which only uses that posteriors are probabilities in $[0, 1]$ and does not depend on η_i), the reputation-for-accuracy difference between any pair of strategies with $a_i = 1$ in one case and $a_i = 0$ in the other satisfies

$$\Delta\text{Acc} \;\geq\; -\kappa(1 - \lambda).$$

Thus the expressive plus reputation-for-accuracy term satisfies

$$\phi + \Delta\text{Acc} \;\geq\; \phi - \kappa(1 - \lambda). \tag{5}$$

Stigma term. From (4), we have $\mu(\Psi^{Thr}) \geq \mu_{\min}^{Thr}$ in every state and equilibrium realization. Hence

$$\min\Big\{M, \frac{1}{\mu(\Psi^{Thr})}\Big\} \;\leq\; \min\Big\{M, \frac{1}{\mu_{\min}^{Thr}}\Big\}.$$

Let

$$S^{Thr} := \min\Big\{M, \frac{1}{\mu_{\min}^{Thr}}\Big\}.$$

Then, for any strategy σ,

$$0 \;\leq\; \mathbb{E}_i\Big[f(\mu(\Psi))P_j(a_i = 1 \mid \Psi^{Thr}, \omega)\Big]_{\sigma} \;\leq\; S^{Thr},$$

and therefore

$$\Delta S = \mathbb{E}_i[\cdot]_{\sigma^{\text{tr}}} - \mathbb{E}_i[\cdot]_{\sigma^{\text{mr}}} \;\leq\; S^{Thr}.$$

Combining (5) with this bound on ΔS, we obtain

$$u_i((1, \eta_H, 0, 0), \sigma^{\text{tr}}) - u_i((1, \eta_H, 0, 0), \sigma^{\text{mr}}) \;\geq\; \big[\phi - \kappa(1-\lambda)\big] - \eta_H \, S^{Thr}.$$

Thus a sufficient condition for the right-hand side to be strictly positive is

$$\eta_H < \frac{\phi - \kappa(1-\lambda)}{S^{Thr}}.$$

By construction $S^{Thr} = \min\{M, 1/\mu^{Thr}_{\min}\} \leq 1/\mu^{Thr}_{\min}$, so

$$\frac{1}{S^{Thr}} \;\geq\; \mu^{Thr}_{\min}, \qquad \Rightarrow \qquad \frac{\phi - \kappa(1-\lambda)}{S^{Thr}} \;\geq\; \mu^{Thr}_{\min}\,\big(\phi - \kappa(1-\lambda)\big).$$

Hence the simpler condition

$$\eta_H < \mu^{Thr}_{\min}\,\big(\phi - \kappa(1-\lambda)\big) =: \bar{\eta}^{(1)}$$

is sufficient to guarantee

$$\eta_H < \frac{\phi - \kappa(1-\lambda)}{S^{Thr}},$$

and therefore

$$u_i((1, \eta_H, 0, 0), \sigma^{\text{tr}}) - u_i((1, \eta_H, 0, 0), \sigma^{\text{mr}}) > 0$$

for every misreporting strategy σ^{mr}.

(c) Conclusion.

Part (a) shows that abstention is strictly dominated by truthful voting with $t_i = 1$, independently of η_H, and part (b) shows that for any $\eta_H \in (0, \bar{\eta}^{(1)})$ truthful voting $a_i = 1$ strictly dominates any misreport $a_i = 0$. Thus, for any $\eta_H \in (0, \bar{\eta}^{(1)})$ the type $(1, \eta_H, 0, 0)$ must participate and choose $a_i = 1$ in every equilibrium.

Sub-lemma B.2. *There exist $\bar{c}, \bar{\pi}, \bar{\phi}, \underline{\eta}, \bar{\eta} > 0$ with $\underline{\eta} < \bar{\eta}$ such that, if $c_H > \bar{c}$, $\pi_H > \bar{\pi}$,*

$\phi > \bar{\phi}$, and $\eta_H \in (\underline{\eta}, \bar{\eta})$, the following holds. In all equilibria under threshold majority voting, agents who vote pick the following thresholds:

$$t_i = \begin{cases} 1 & \textit{if } c_i = 0 \wedge \pi_i = \pi_H, \\ \lambda & \textit{if } c_i = 0 \wedge \pi_i = 0 \wedge \eta_i = 0, \\ 1 - \lambda & \textit{if } c_i = 0 \wedge \pi_i = 0 \wedge \eta_i = \eta_H \wedge s_i = 0, \\ 1 & \textit{if } c_i = 0 \wedge \pi_i = 0 \wedge \eta_i = \eta_H \wedge s_i = 1. \end{cases}$$

Proof. We characterize threshold choices for each type class. Step 2.A first establishes the aggregate vote shares implied by truthful voting, which pin down the canonical thresholds. We then analyze threshold choices for low-stigma types (Step 2.B) and high-stigma types by signal (Steps 2.C and 2.D).

Step 2.A: Aggregate vote shares.

Claim. Assume our restrictions on parameters and suppose Sublemma B.1 hold. Then, in any equilibrium of the threshold majority voting mechanism, the equilibrium share $\mu(\Psi^{\text{Thr}})$ of votes for the controversial option $a = 1$ among non-abstainers satisfies

$$\mu(\Psi^{\text{Thr}}) = \begin{cases} 1 - \lambda & \text{if } \omega = 0, \\ \lambda & \text{if } \omega = 1. \end{cases}$$

Proof. By Sublemma B.1, all agents with $c_i = 0$ vote truthfully ($a_i = s_i$) while high-cost types abstain. Threshold choices affect only revelation, not underlying shares.

By the signal structure, for each state $\omega \in \{0, 1\}$ we have

$$\Pr(s_i = \omega \mid \omega) = \lambda, \qquad \Pr(s_i = 1 - \omega \mid \omega) = 1 - \lambda.$$

Since c_i is independent of (s_i, ω) and truthful voting implies $a_i = s_i$,

$$\mu(\Psi^{\mathrm{Thr}} \mid \omega) = \Pr(s_i = 1 \mid \omega) = \begin{cases} 1 - \lambda & \text{if } \omega = 0, \\ \lambda & \text{if } \omega = 1. \end{cases}$$

Step 2.B: Threshold choice of low-cost, low-privacy, low-stigma types.

Low-stigma types have nothing to hide—they simply want to maximize reputation-for-accuracy. The intermediate threshold $t_i = \lambda$ achieves this: they are revealed (and receive full reputation-for-accuracy credit) when their vote matches the realized state, and remain undisclosed (receiving partial credit from the undisclosed pool) otherwise.

Claim. Assume our restrictions on parameters and suppose Sublemma B.1 hold. Consider any agent of type

$$\boldsymbol{x}_i = (s_i, \eta_i, \pi_i, c_i) = (s_i, 0, 0, 0), \quad s_i \in \{0, 1\}.$$

In any equilibrium of the threshold majority voting mechanism, such a type votes truthfully (by Sublemma B.1) and chooses threshold $t_i = \lambda$.

Proof. By Sublemma B.1, $a_i = s_i$. By Step 2.A,

$$\mu_1(0) = 1 - \lambda, \quad \mu_1(1) = \lambda, \qquad \mu_0(0) = \lambda, \quad \mu_0(1) = 1 - \lambda,$$

so $\mu_a^{\min} = 1 - \lambda$ and $\mu_a^{\max} = \lambda$ for each $a \in \{0, 1\}$. By Lemma 3, the canonical thresholds are $t_{s_i}^{\mathrm{low}} = 1 - \lambda$, $t_{s_i}^{\mathrm{int}} = \lambda$, and $t_{s_i}^{\mathrm{high}} = 1$.

For type $(s_i, 0, 0, 0)$, only the reputation-for-accuracy term depends on t_i.

By Lemma 4: $P_j(a_i = \omega \mid L^{\omega}, \omega) = 1$, $P_j(a_i = \omega \mid L^{1-\omega}, \omega) = 0$, and $q_\omega := P_j(a_i = \omega \mid L^u, \omega) \in (0, 1)$.

Let $\omega_{\mathrm{good}} := s_i$ (correct state) and $\omega_{\mathrm{bad}} := 1 - s_i$ (incorrect state), with $\Pr(\omega_{\mathrm{good}} \mid s_i) = \lambda$.

Low threshold $t_{s_i}^{\mathrm{low}} = 1 - \lambda$. Revealed in both states with label L^{s_i}: $U^{\mathrm{low}} = \kappa[\lambda \cdot 1 + (1 - \lambda) \cdot 0] = \kappa\lambda$.

Intermediate threshold $t_{s_i}^{\mathrm{int}} = \lambda$. Revealed in good state (L^{s_i}), undisclosed in bad state (L^u): $U^{\mathrm{int}} = \kappa[\lambda \cdot 1 + (1 - \lambda) \cdot q_{\omega_{\mathrm{bad}}}] > \kappa\lambda = U^{\mathrm{low}}$.

High threshold $t_{s_i}^{\text{high}} = 1$. Never revealed ($L^u$ in both states): $U^{\text{high}} = \kappa[\lambda \cdot q_{\omega_{\text{good}}} + (1-\lambda) \cdot q_{\omega_{\text{bad}}}]$. Since $q_{\omega_{\text{good}}} < 1$, $U^{\text{int}} - U^{\text{high}} = \kappa\lambda(1 - q_{\omega_{\text{good}}}) > 0$.

Comparison. $U^{\text{int}} > U^{\text{low}}$ and $U^{\text{int}} > U^{\text{high}}$, so the intermediate threshold strictly maximizes utility. By the tie-breaking convention, the unique optimal threshold is $t_i = \lambda$. □

Step 2.C: Threshold choice of high-stigma types with $s_i = 0$.

Claim. Assume our restrictions on parameters and suppose Sublemma B.1 hold. Define

$$\eta_0 := \frac{\kappa}{1-\lambda}.$$

Then, for every $\eta_H > \eta_0$, in any equilibrium of the threshold majority voting mechanism, any agent of type

$$\boldsymbol{x}_i = (s_i, \eta_i, \pi_i, c_i) = (0, \eta_H, 0, 0)$$

votes truthfully ($a_i = 0$) and chooses the low canonical threshold

$$t_i = t_0^{\text{low}} = 1 - \lambda.$$

Proof. By Sublemma B.1, $a_i = 0$. By Step 2.A, $\mu_0^{\min} = 1 - \lambda$ and $\mu_0^{\max} = \lambda$, giving canonical thresholds $t_0^{\text{low}} = 1 - \lambda$, $t_0^{\text{int}} = \lambda$, and $t_0^{\text{high}} = 1$.

Let

$$q_1^1 := P_j(a_i = 1 \mid L^u, \omega = 1) \in (0,1), \qquad q_0^1 := P_j(a_i = 1 \mid L^u, \omega = 0) \in (0,1),$$

and similarly $q_0^0 := P_j(a_i = 0 \mid L^u, \omega = 0) \in (0,1)$. We compare the three canonical thresholds.

(i) t_0^{high} is strictly dominated by t_0^{int}.

Under $t_0^{\text{high}} = 1$, the agent is never revealed and always has label $L_i = L^u$ in both states.

Under $t_0^{\text{int}} = \lambda$, she is revealed as L^0 in state $\omega = 0$ and undisclosed L^u in state $\omega = 1$.

reputation-for-accuracy. In $\omega = 0$, under t_0^{int} the label is L^0 and $P_j(a_i = \omega \mid L^0, 0) = P_j(a_i = 0 \mid L^0, 0) = 1$, while under t_0^{high} the label is L^u and $P_j(a_i = \omega \mid L^u, 0) = q_0^0 \in (0,1)$. In $\omega = 1$, both thresholds yield L^u. Hence reputation-for-accuracy is strictly higher under

t_0^{int}.

Stigma. In state $\omega = 0$, $\mu(\Psi^{\text{Thr}}) = 1 - \lambda$ so $1/\mu(\Psi^{\text{Thr}}) = 1/(1-\lambda)$; in state $\omega = 1$, $\mu(\Psi^{\text{Thr}}) = \lambda$ so $1/\mu(\Psi^{\text{Thr}}) = 1/\lambda$. Under t_0^{int}:

- for $\omega = 0$ the label is L^0, so $P_j(a_i = 1 \mid L^0, 0) = 0$ and the stigma contribution is 0;

- for $\omega = 1$ the label is L^u, so the contribution is

$$\frac{1}{\lambda} P_j(a_i = 1 \mid L^u, 1) = \frac{1}{\lambda} q_1^1.$$

Under t_0^{high} the agent is L^u in both states, so the contribution is

$$\frac{1}{1-\lambda} q_0^1 \quad \text{in } \omega = 0, \qquad \frac{1}{\lambda} q_1^1 \quad \text{in } \omega = 1.$$

Conditioning on $s_i = 0$ (so that $\Pr(\omega = 0 \mid s_i = 0) = \lambda$ and $\Pr(\omega = 1 \mid s_i = 0) = 1 - \lambda$), we obtain

$$S^{\text{int}} := \mathbb{E}_i \left[\frac{1}{\mu(\Psi^{\text{Thr}})} P_j(a_i = 1 \mid \Psi^{\text{Thr}}, \omega) \right]_{t_0^{\text{int}}} = (1-\lambda) \cdot \frac{1}{\lambda} q_1^1 = \frac{1-\lambda}{\lambda} q_1^1,$$

$$S^{\text{high}} := \mathbb{E}_i \left[\frac{1}{\mu(\Psi^{\text{Thr}})} P_j(a_i = 1 \mid \Psi^{\text{Thr}}, \omega) \right]_{t_0^{\text{high}}} = \lambda \cdot \frac{1}{1-\lambda} q_0^1 + (1-\lambda) \cdot \frac{1}{\lambda} q_1^1.$$

Since $q_0^1 > 0$, we have $S^{\text{high}} > S^{\text{int}}$, so stigma is strictly more negative under t_0^{high} than under t_0^{int}.

Thus t_0^{high} yields both lower reputation-for-accuracy and more negative stigma; it is strictly dominated by t_0^{int} and can be ignored.

(ii) Explicit comparison of t_0^{low} and t_0^{int}.

We now compare $t_0^{\text{low}} = 1 - \lambda$ and $t_0^{\text{int}} = \lambda$.

Under t_0^{low}, the agent is revealed as L^0 in both states. Hence

$$\text{Acc}^{\text{low}} = \kappa\lambda$$

and, since $P_j(a_i = 1 \mid L^0, \omega) = 0$ in both states, the stigma term is identically zero:

$$\mathrm{Stig}^{\mathrm{low}} = 0.$$

Under t_0^{int}, the agent is L^0 in $\omega = 0$ and L^u in $\omega = 1$. As above,

$$\mathrm{Acc}^{\mathrm{int}} = \kappa\big[\lambda \cdot 1 + (1-\lambda) P_j(a_i = 1 \mid L^u, 1)\big] = \kappa\big[\lambda + (1-\lambda) q_1^1\big],$$

so

$$\mathrm{Acc}^{\mathrm{low}} - \mathrm{Acc}^{\mathrm{int}} = -\kappa(1-\lambda) q_1^1.$$

For the stigma term, we already computed

$$S^{\mathrm{int}} := \mathbb{E}_i\left[\frac{1}{\mu(\Psi^{\mathrm{Thr}})} P_j(a_i = 1 \mid \Psi^{\mathrm{Thr}}, \omega)\right]_{t_0^{\mathrm{int}}} = \frac{1-\lambda}{\lambda} q_1^1,$$

so

$$\mathrm{Stig}^{\mathrm{low}} - \mathrm{Stig}^{\mathrm{int}} = 0 - \big[-\eta_H S^{\mathrm{int}}\big] = \eta_H \, \frac{1-\lambda}{\lambda} q_1^1.$$

The total utility difference is therefore

$$u_i((0, \eta_H, 0, 0), 0, t_0^{\mathrm{low}}) - u_i((0, \eta_H, 0, 0), 0, t_0^{\mathrm{int}}) = (1-\lambda) q_1^1 \Big[\frac{\eta_H}{\lambda} - \kappa\Big].$$

Since $q_1^1 > 0$ and $1 - \lambda > 0$, the sign is the sign of $\frac{\eta_H}{\lambda} - \kappa$.

We consider $\eta_H > \eta_0 = \kappa/(1-\lambda)$ and $\lambda \in (1/2, 1)$. Hence $\eta_0 > \kappa\lambda$, so for every $\eta_H > \eta_0$ we have

$$\frac{\eta_H}{\lambda} > \frac{\kappa\lambda}{\lambda} = \kappa,$$

and thus

$$u_i((0, \eta_H, 0, 0), 0, t_0^{\mathrm{low}}) - u_i((0, \eta_H, 0, 0), 0, t_0^{\mathrm{int}}) > 0.$$

Combining (i) and (ii), we conclude that for every $\eta_H > \eta_0$ the unique optimal canonical threshold on side $a_i = 0$ is $t_0^{\mathrm{low}} = 1 - \lambda$. By Lemma 3 and the tie-breaking rule, any best response of type $(0, \eta_H, 0, 0)$ can be represented by $t_i = 1 - \lambda$.

Note that the condition $\eta_H > \eta_0 := \kappa/(1-\lambda)$ is a sufficient—but not necessary—lower bound for t_0^{low} to dominate t_0^{int}. The minimal requirement obtained from the comparison above is $\eta_H > \kappa\lambda$. Since $\lambda > 1/2$, we have $\kappa/(1-\lambda) > \kappa\lambda$, so imposing $\eta_H > \eta_0$ is strictly stronger. We adopt this bound because it simplifies the algebra in the subsequent parameter-compatibility analysis, where the term $\kappa/(1-\lambda)$ naturally appears alongside other expressions involving $1-\lambda$. □

Step 2.D: Threshold choice of high-stigma types with $s_i = 1$.

High-stigma agents voting for the controversial option $a = 1$ face a genuine tradeoff: revelation brings reputation-for-accuracy benefits but also stigma costs. When stigma sensitivity is sufficiently high, these agents prefer to never reveal, staying in the undisclosed pool alongside abstainers and others whose votes remain hidden.

Claim. Assume our restrictions on parameters and suppose Sublemma B.1 hold, and let $\eta_0 = \kappa/(1-\lambda)$ as in Step 2.C. Then, for every $\eta_H > \eta_0$, in any equilibrium of the threshold majority voting mechanism, any agent of type

$$\boldsymbol{x}_i = (s_i, \eta_i, \pi_i, c_i) = (1, \eta_H, 0, 0)$$

votes truthfully ($a_i = 1$) and chooses the high canonical threshold

$$t_i = t_1^{\text{high}} = 1.$$

Proof. By Sublemma B.1, $a_i = 1$. By Step 2.A, $\mu_1^{\min} = 1-\lambda$ and $\mu_1^{\max} = \lambda$, giving canonical thresholds $t_1^{\text{low}} = 1-\lambda$, $t_1^{\text{int}} = \lambda$, and $t_1^{\text{high}} = 1$. Let

$$q_1^1 := P_j(a_i = 1 \mid L^u, \omega = 1) \in (0,1),$$
$$q_0^0 := P_j(a_i = 0 \mid L^u, \omega = 0) \in (0,1),$$
$$q_0^1 := P_j(a_i = 1 \mid L^u, \omega = 0) \in (0,1).$$

(i) t_1^{low} is strictly dominated by t_1^{int}.

Under $t_1^{\text{low}} = 1-\lambda$, the agent is revealed as L^1 in both states. Under $t_1^{\text{int}} = \lambda$, she is L^u

in $\omega = 0$ and L^1 in $\omega = 1$.

reputation-for-accuracy. Conditioning on $s_i = 1$ (so that $\Pr(\omega = 1 \mid s_i = 1) = \lambda$ and $\Pr(\omega = 0 \mid s_i = 1) = 1 - \lambda$), we have:

$$\mathrm{Acc}^{\mathrm{low}} = \kappa\big[\lambda \cdot 1 + (1-\lambda) \cdot 0\big] = \kappa\lambda,$$

$$\mathrm{Acc}^{\mathrm{int}} = \kappa\big[\lambda \cdot 1 + (1-\lambda) P_j(a_i = 0 \mid L^u, 0)\big] = \kappa\big[\lambda + (1-\lambda) q_0^0\big] > \kappa\lambda.$$

Stigma. Under t_1^{low}, the label is L^1 in both states, and $P_j(a_i = 1 \mid L^1, \omega) = 1$ for $\omega = 0, 1$. Using

$$\mu(\Psi^{\mathrm{Thr}}) = \begin{cases} 1 - \lambda & \text{if } \omega = 0, \\ \lambda & \text{if } \omega = 1, \end{cases}$$

we get

$$S^{\mathrm{low}} := \mathbb{E}_i\left[\frac{1}{\mu(\Psi^{\mathrm{Thr}})} P_j(a_i = 1 \mid \Psi^{\mathrm{Thr}}, \omega)\right]_{t_1^{\mathrm{low}}} = (1-\lambda) \cdot \frac{1}{1-\lambda} \cdot 1 + \lambda \cdot \frac{1}{\lambda} \cdot 1 = 2.$$

Under t_1^{int}, the agent is L^u in $\omega = 0$ and L^1 in $\omega = 1$, so

$$S^{\mathrm{int}} = (1-\lambda) \cdot \frac{1}{1-\lambda} q_0^1 + \lambda \cdot \frac{1}{\lambda} \cdot 1 = q_0^1 + 1.$$

Since $q_0^1 \in (0, 1)$, we have $S^{\mathrm{int}} < S^{\mathrm{low}}$, so stigma is strictly less negative under t_1^{int}:

$$\mathrm{Stig}^{\mathrm{int}} - \mathrm{Stig}^{\mathrm{low}} = -\eta_H S^{\mathrm{int}} + \eta_H S^{\mathrm{low}} = \eta_H (1 - q_0^1) > 0.$$

Both reputation-for-accuracy and stigma strictly favour t_1^{int} over t_1^{low}, so t_1^{low} is strictly dominated and can be ignored.

(ii) Explicit comparison of t_1^{int} and t_1^{high}.

We now compare $t_1^{\mathrm{int}} = \lambda$ and $t_1^{\mathrm{high}} = 1$.

Under t_1^{int}, the agent is L^u in $\omega = 0$ and L^1 in $\omega = 1$. Under t_1^{high}, she is always L^u.

reputation-for-accuracy. As above,

$$\mathrm{Acc}^{\mathrm{int}} = \kappa\big[\lambda \cdot 1 + (1-\lambda)q_0^0\big], \qquad \mathrm{Acc}^{\mathrm{high}} = \kappa\big[\lambda q_1^1 + (1-\lambda)q_0^0\big],$$

so

$$\mathrm{Acc}^{\mathrm{int}} - \mathrm{Acc}^{\mathrm{high}} = \kappa\lambda(1-q_1^1) > 0.$$

Stigma. Under t_1^{int} we already have

$$S^{\mathrm{int}} = q_0^1 + 1.$$

Under t_1^{high}, the label is L^u in both states, so

$$S^{\mathrm{high}} = (1-\lambda)\cdot\frac{1}{1-\lambda}q_0^1 + \lambda\cdot\frac{1}{\lambda}q_1^1 = q_0^1 + q_1^1.$$

Thus

$$S^{\mathrm{int}} - S^{\mathrm{high}} = 1 - q_1^1 > 0,$$

and the stigma components satisfy

$$\mathrm{Stig}^{\mathrm{int}} - \mathrm{Stig}^{\mathrm{high}} = -\eta_H\big(S^{\mathrm{int}} - S^{\mathrm{high}}\big) = -\eta_H(1-q_1^1).$$

Total difference. The total utility difference between t_1^{int} and t_1^{high} is

$$u_i((1,\eta_H,0,0),1,t_1^{\mathrm{int}}) - u_i((1,\eta_H,0,0),1,t_1^{\mathrm{high}}) = \kappa\lambda(1-q_1^1) - \eta_H(1-q_1^1) = (1-q_1^1)\big[\kappa\lambda - \eta_H\big].$$

Since $1-q_1^1 > 0$, the sign is the sign of $\kappa\lambda - \eta_H$. For $\eta_H > \eta_0 = \kappa/(1-\lambda)$ and $\lambda \in (1/2, 1)$ we have $\eta_H > \kappa\lambda$, so $\kappa\lambda - \eta_H < 0$, and hence

$$u_i((1,\eta_H,0,0),1,t_1^{\mathrm{int}}) - u_i((1,\eta_H,0,0),1,t_1^{\mathrm{high}}) < 0.$$

Thus, whenever $\eta_H > \eta_0$, $t_1^{\mathrm{high}} = 1$ strictly dominates t_1^{int}.

Combining (i) and (ii), we conclude that for every $\eta_H > \eta_0$ the unique optimal canonical threshold on side $a_i = 1$ is $t_1^{\text{high}} = 1$. By Lemma 3 and the tie-breaking rule, any best response of type $(1, \eta_H, 0, 0)$ can be represented by $t_i = 1$. □

Step 2.E: Conclusion of the proof of Sublemma B.2.

We now combine the previous steps to establish the sublemma.

By Sublemma B.1, and under our restrictions on parameters, there exists a constant $\bar{\eta}^{(1)} > 0$ such that, for every $\eta_H < \bar{\eta}^{(1)}$, in all equilibria of the threshold mechanism: (i) all high-cost types with $c_i = c_H$ abstain; (ii) all low-cost types with $c_i = 0$ participate and vote truthfully, $a_i = s_i$.

Given truthful voting, Step 2.A shows that, in any equilibrium, the share $\mu(\Psi^{\text{Thr}})$ of votes for the controversial option $a = 1$ among non-abstainers is

$$\mu(\Psi^{\text{Thr}}) = \begin{cases} 1 - \lambda & \text{if } \omega = 0, \\ \\ \lambda & \text{if } \omega = 1. \end{cases}$$

Next, we collect the threshold choices type by type:

- *Low-cost, high-privacy types* $(c_i = 0, \pi_i = \pi_H)$: by the analysis in Sublemma B.1 (its Step B), such types vote truthfully and strictly prefer the "never reveal" option $t_i = 1$ for any η_H. Their threshold choice does not depend on η_H.

- *Low-cost, low-privacy, low-stigma types* $(c_i = 0, \pi_i = 0, \eta_i = 0)$: Step 2.B shows that, for any η_H and any equilibrium, such types vote truthfully and choose the intermediate canonical threshold $t_i = \lambda$.

- *Low-cost, low-privacy, high-stigma types with* $s_i = 0$: Step 2.C shows that, for these types (who vote $a_i = 0$), the unique best-response threshold is $t_i = 1 - \lambda$ whenever $\eta_H > \eta_0 = \kappa/(1 - \lambda)$.

- *Low-cost, low-privacy, high-stigma types with* $s_i = 1$: Step 2.D shows that, for these types (who vote $a_i = 1$), the unique best-response threshold is $t_i = 1$ whenever $\eta_H > \eta_0 = \kappa/(1 - \lambda)$.

Hence, for any parameter η_H satisfying

$$\eta_H > \eta_0 = \frac{\kappa}{1-\lambda} \quad \text{and} \quad \eta_H < \bar{\eta}^{(1)},$$

the threshold best responses of all low-cost types are uniquely determined in any equilibrium and coincide with

$$t_i = \begin{cases} 1 & \text{if } c_i = 0 \land \pi_i = \pi_H, \\ \lambda & \text{if } c_i = 0 \land \pi_i = 0 \land \eta_i = 0, \\ 1-\lambda & \text{if } c_i = 0 \land \pi_i = 0 \land \eta_i = \eta_H \land s_i = 0, \\ 1 & \text{if } c_i = 0 \land \pi_i = 0 \land \eta_i = \eta_H \land s_i = 1. \end{cases}$$

The next section of the proof shows that $\eta_0 < \bar{\eta}^{(1)}$ under our parameter restrictions, so the interval $(\eta_0, \bar{\eta}^{(1)})$ is non-empty.

□

Parameter consistency of the bounds on η_H.

Our analysis of the three voting mechanisms imposes several restrictions on the stigma parameter η_H. The key quantities are: $\eta_0 := \kappa/(1-\lambda)$ (threshold for high-stigma behavior, from Steps 2.C–2.D), q^{tr} (mass of always-truthful types, from Step 1.C), and $\mu_{\min}^{Thr} := q^{\text{tr}}(1-\lambda)$ (lower bound on controversial vote share).

- From Sublemma B.1 (truthful voting under the threshold mechanism), we require an *upper* bound

  $$\eta_H < \bar{\eta}^{(1)} := \mu_{\min}^{Thr}\big(\phi - \kappa(1-\lambda)\big),$$

  so that expressive plus reputation-for-accuracy concerns dominate the maximum possible stigma gain from misreporting.

- From Sublemma B.2 (threshold choice of high-stigma types under the threshold mechanism), we require a *lower* bound

$$\eta_H > \eta_0 := \frac{\kappa}{1-\lambda},$$

in order to ensure that high-stigma types strictly prefer the extreme thresholds (low for $s_i = 0$, high for $s_i = 1$) over the intermediate threshold.

- From Lemma 2 (public majority voting), we obtain another *lower* bound

$$\eta_H > \underline{\eta}^{Pub},$$

where $\underline{\eta}^{Pub} > 0$ is the minimal stigma level that makes it optimal for high-stigma, low-privacy, low-cost types with $s_i = 1$ to vote for the non-controversial option $a_i = 0$ rather than $a_i = 1$.

For these requirements to be compatible, we need the intersection

$$\big(\eta_0, \bar{\eta}^{(1)}\big) \ \cap\ \big(\underline{\eta}^{Pub}, \infty\big) = \big(\max\{\eta_0, \underline{\eta}^{Pub}\},\ \bar{\eta}^{(1)}\big)$$

to be nonempty. This is equivalent to

$$\max\{\eta_0, \underline{\eta}^{Pub}\} < \bar{\eta}^{(1)}. \tag{6}$$

We now express (6) as a simple restriction on the primitives (ϕ, κ, λ), given the lower bound $\mu_{\min}^{Thr} > 0$ defined in Sublemma B.1.

First, the inequality $\eta_0 < \bar{\eta}^{(1)}$ is

$$\frac{\kappa}{1-\lambda} < \mu_{\min}^{Thr}\big(\phi - \kappa(1-\lambda)\big),$$

or, rearranging,

$$\phi > \frac{\kappa}{(1-\lambda)\mu_{\min}^{Thr}} + \kappa(1-\lambda).$$

Thus

$$\phi > \kappa \left(\frac{1}{(1-\lambda)\mu_{\min}^{Thr}} + 1 - \lambda \right) \tag{7}$$

is necessary and sufficient for the interval $(\eta_0, \bar{\eta}^{(1)})$ to be nonempty.

Second, we require $\underline{\eta}^{Pub} < \bar{\eta}^{(1)}$. In the proof of Lemma 2 we obtained

$$\underline{\eta}^{Pub} \ = \ \frac{\phi + \kappa(2\lambda - 1)}{S^{Pub}},$$

for some constant $S^{Pub} > 0$. From the public-voting equilibrium,

$$\mu_1(0) = p_{\eta_L}(1-\lambda), \qquad \mu_1(1) = p_{\eta_L}\lambda.$$

When the cap in $f(\cdot)$ does not bind (i.e. when $M > \frac{1}{p_{\eta_L}(1-\lambda)}$),

$$S^{Pub} = \frac{2}{p_{\eta_L}}.$$

On the threshold side, Sublemma B.1 implies

$$\mu(\Psi^{Thr} \mid \omega) \ \geq \ \mu_{\min}^{Thr} := q^{\text{tr}}(1-\lambda),$$

where $q^{\text{tr}} := p_{c_L}p_{\pi_H} + p_{c_L}p_{\pi_L}p_{\eta_L}$ is the mass of low-cost types who always vote truthfully. It follows that

$$\mu_{\min}^{Thr} S^{Pub} = \frac{2q^{\text{tr}}(1-\lambda)}{p_{\eta_L}}.$$

The condition $\mu_{\min}^{Thr} S^{Pub} > 1$ is equivalent to

$$q^{\text{tr}} > \frac{p_{\eta_L}}{2(1-\lambda)}.$$

We impose this lower bound on q^{tr}.

For notational convenience, let

$$A \ := \ \mu_{\min}^{Thr} S^{Pub} \ > \ 1.$$

Then the inequality $\underline{\eta}^{Pub} < \bar{\eta}^{(1)}$ is equivalent to

$$\begin{aligned} & \frac{\phi + \kappa(2\lambda - 1)}{S^{Pub}} < \mu_{\min}^{Thr}\big(\phi - \kappa(1-\lambda)\big) \\ \iff \quad & \phi + \kappa(2\lambda - 1) < A\big(\phi - \kappa(1-\lambda)\big) \\ \iff \quad & (A-1)\phi > \kappa\big(A(1-\lambda) + (2\lambda - 1)\big). \end{aligned}$$

Since $A > 1$ and $2\lambda - 1 > 0$, the right-hand side is strictly positive. Hence there exists a finite constant

$$\bar{\phi}^{(2)} \;:=\; \frac{\kappa\big(A(1-\lambda) + (2\lambda - 1)\big)}{A-1} \tag{8}$$

such that, for all $\phi > \bar{\phi}^{(2)}$, we have

$$\underline{\eta}^{Pub} < \bar{\eta}^{(1)}.$$

Combining (7) and (8), we can define

$$\bar{\phi} \;:=\; \max\left\{\kappa\left(\frac{1}{(1-\lambda)\mu_{\min}^{Thr}} + 1 - \lambda\right),\; \bar{\phi}^{(2)}\right\},$$

so that for every $\phi > \bar{\phi}$ both inequalities $\eta_0 < \bar{\eta}^{(1)}$ and $\underline{\eta}^{Pub} < \bar{\eta}^{(1)}$ hold. In particular, for all $\phi > \bar{\phi}$, the compatibility condition (6) is satisfied.

Implications for the main interval $(\underline{\eta}, \bar{\eta})$.

Sublemma B.2 shows that, provided $\eta_0 < \bar{\eta}^{(1)}$, there exist constants $\underline{\eta}^{Thr}, \bar{\eta}^{Thr}$ with

$$\eta_0 < \underline{\eta}^{Thr} < \bar{\eta}^{Thr} \leq \bar{\eta}^{(1)}$$

such that the equilibrium characterization of the threshold mechanism holds for all $\eta_H \in (\underline{\eta}^{Thr}, \bar{\eta}^{Thr})$ (for any type distribution satisfying the lower bound on q^{tr} used above).

Combining this with the bound from public majority voting, we can define

$$\underline{\eta} \;:=\; \max\{\underline{\eta}^{Pub}, \underline{\eta}^{Thr}\}, \qquad \bar{\eta} \;:=\; \bar{\eta}^{Thr}.$$

Under our restrictions on the primitives and for all $\phi > \bar{\phi}$, we have

$$\underline{\eta} \;\geq\; \max\{\eta_0, \underline{\eta}^{Pub}\} \quad \text{and} \quad \bar{\eta} \;\leq\; \bar{\eta}^{(1)},$$

and the compatibility condition (6) implies

$$\underline{\eta} < \bar{\eta}.$$

Hence the interval $(\underline{\eta}, \bar{\eta})$ is nonempty.

By construction, for every $\eta_H \in (\underline{\eta}, \bar{\eta})$ and $\phi > \bar{\phi}$:

- the public majority-voting equilibrium described in Lemma 2 exists (because $\eta_H > \underline{\eta}^{Pub}$);
- the threshold majority-voting equilibrium described in Lemma 5 exists and is unique (because $\eta_H \in (\underline{\eta}^{Thr}, \bar{\eta}^{Thr})$ and the lower bound on q^{tr} holds);
- the anonymous majority-voting equilibrium described in Lemma 1 exists and is unique.

Therefore, the constants $\underline{\eta}, \bar{\eta} > 0$ appearing in Proposition 2 can be chosen exactly as above, and $(\underline{\eta}, \bar{\eta})$ is a nonempty interval on which all three equilibrium characterizations (anonymous, public, and threshold) hold simultaneously. This delivers the parameter-consistency statement claimed in Proposition 2.

□

B.3 Summary: The Three Channels

The equilibrium characterized in Proposition 2 illustrates how threshold majority voting operates through the three channels described in the main text. The *privacy channel* appears in Step 1.B: high-privacy types participate and vote truthfully because they can choose $t_i = 1$ and avoid any disclosure cost. The *epistemic channel* is visible in Step 2.B: low-stigma types choose the intermediate threshold $t_i = \lambda$, which reveals their vote exactly when the realized state matches their signal—that is, when their vote is ex post correct. The *Safety-in-numbers channel* appears in Steps 2.C and 2.D together with the lower bound construction in Step 1.C: high-stigma types with $s_i = 1$ can vote truthfully while choosing a high threshold $t_i = 1$, which keeps them in the undisclosed pool L^u alongside abstainers and other never-reveal voters ("social cover"). Because the stigma term falls when the controversial option attracts more support (via $f(\mu(\Psi))$), the expected stigma from being suspected of voting 1 is mitigated when $\mu(\Psi)$ is larger, thus making the "social cover" effectively more appealing.

C Survey Details and Sample Characteristics

C.1 Pre-registration

We pre-registered our design, hypotheses, and analysis plan on the AEA RCT Registry (AEARCTR-16968) prior to data collection. The pre-analysis plan specified comparisons of abstention rates (H1) and expression of controversial views (H2) across treatments, with Public expected to increase abstention and suppress controversial expression relative to both Private and Threshold. Vote revelation rates in the Threshold treatment (H4) were pre-registered as a secondary outcome. All primary analyses reported in the main text follow the pre-registered specifications.

Two outcomes were not pre-registered: the uncontroversial vote share among non-abstainers (H3), which conditions on participation rather than measuring expression among all participants as in the PAP; and the distributional comparison of threshold choices by vote direction (H5).

The final sample (N = 298) is smaller than originally projected. The pre-analysis plan anticipated approximately 1,200 participants based on XLab's initial estimate of their subject pool; the available pool of eligible participants proved smaller than projected. The pre-registered sequential design specified expanding to a second university if preconditions were not met at Stage 1. Because the preconditions were satisfied with the UC Berkeley sample, we did not proceed to Stage 2.

Heterogeneity analyses by political ideology, gender, and engagement were pre-specified as exploratory. Robustness checks excluding fast respondents and participants who revised their threshold choice were not pre-registered.

C.2 Demographic Summary Statistic

Table C1 presents demographic summary statistics for the analytic sample. The analytical sample consists of 298 undergraduates enrolled at UC Berkeley, recruited via the Expermential Social Science Laboratory (XLab) in October–November 2025. The sample has a median age of 20 years, with 68.5% identifying as female and 33.2% as non-heterosexual.

The sample is predominantly liberal (75.5%), reflecting UC Berkeley's political composition, with a liberal-to-conservative ratio of 10.2:1. The largest racial/ethnic groups are Asian or Asian-American (65 %), White (23 %), and Hispanic or Latino (15 %). Fields of study are diverse, with the largest shares in Engineering & Computer Science (29 %), Life Sciences & Medicine (22 %), and Social Sciences (18 %).

C.3 Randomization balance

Table C2 confirms that randomization produced balanced groups across baseline covariates.

C.4 Data Quality

We received 358 survey submissions. After excluding participants who did not consent or pass eligibility screening (UC Berkeley undergraduate enrollment), failed the attention check, or submitted duplicate responses, 328 were randomized to treatment. Of these, 30 attrited before completion, yielding a final sample of N=298.

Table C3 summarizes data quality metrics. Median completion time was 6.3 minutes. Most participants passed comprehension checks on the first attempt: 61% passed vote comprehension checks, and 79% of Threshold participants passed threshold-specific comprehension checks on the first try.

Appendix Table C1: Demographic Summary Statistics

	Mean	SD
	(1)	(2)
Demographics		
Age	20.21	2.66
Graduation year	2027.15	1.25
Female (%)	68.46	46.55
Other/Non-binary gender (%)	4.36	20.46
Non-heterosexual (%)	33.22	47.18
Race/Ethnicity		
White (%)	23.15	42.25
Black or African-American (%)	2.68	16.19
Hispanic or Latino (%)	15.10	35.87
Asian or Asian-American (%)	65.10	47.75
Other race/ethnicity (%)	4.70	21.20
Politics		
Conservative ideology (%)	7.38	26.19
Liberal ideology (%)	75.50	43.08
Democrat (including leaners) (%)	89.93	30.14
Republican (including leaners) (%)	10.07	30.14
Field of Study		
Arts and Humanities (%)	7.05	25.64
Social Sciences (%)	17.79	38.30
Business and Economics (%)	13.09	33.78
Engineering and Computer Science (%)	28.86	45.39
Life Sciences and Medicine (%)	22.15	41.59
Other field (%)	11.07	31.43

Note: This table presents summary statistics for participants who passed the attention check and completed the study (N = 298). Column (1): sample mean. Column (2): standard deviation. Variables labeled with (%) are binary indicators expressed as percentages. Age is in years; graduation year is expected year of degree completion. Gender and sexual orientation are self-reported. Race/ethnicity categories are not mutually exclusive. Political ideology is measured on a 7-point scale from "Very Conservative" (1) to "Very Liberal" (7); liberal includes responses 5–7, conservative includes 1–3 (4 = moderate). Party identification includes those who identify with or lean toward each party. Field of study reflects primary academic major.

Appendix Table C2: Baseline Balance Across Treatment Arms

	Means			p-values			
	Pri	Pub	Thr	Pub –Pri	Thr –Pri	Thr –Pub	Joint
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Demographics							
Age	20.05	20.09	20.56	0.907	0.163	0.275	0.340
Graduation year	2027	2027	2027	0.410	0.626	0.194	0.427
Female (%)	66.98	71.43	66.67	0.487	0.963	0.479	0.719
Non-binary/Other (%)	1.89	4.76	6.90	0.246	0.083	0.529	0.232
Non-heterosexual (%)	33.02	32.38	34.48	0.922	0.832	0.760	0.953
Race/Ethnicity							
White (%)	22.64	23.81	22.99	0.842	0.955	0.894	0.979
Hispanic or Latino (%)	14.15	18.10	12.64	0.438	0.762	0.303	0.546
Asian or Asian-Amer. (%)	65.09	60.95	70.11	0.535	0.462	0.187	0.418
Politics							
Ideology (7-point scale)	5.22	5.33	5.30	0.442	0.607	0.837	0.740
Conservative ideology (%)	6.60	7.62	8.05	0.775	0.703	0.913	0.924
Liberal ideology (%)	75.47	77.14	73.56	0.777	0.763	0.568	0.849
Democrat (+leaners) (%)	87.74	94.29	87.36	0.097	0.937	0.093	0.184
Republican (+leaners) (%)	12.26	5.71	12.64	0.097	0.937	0.093	0.184
Field of Study							
Arts and Humanities (%)	8.49	7.62	4.60	0.817	0.285	0.392	0.555
Social Sciences (%)	15.09	18.10	20.69	0.560	0.312	0.652	0.599
Business & Econ. (%)	16.04	10.48	12.64	0.236	0.508	0.641	0.486
Engineering and CS (%)	25.47	29.52	32.18	0.512	0.307	0.693	0.584
Life Sci. & Med. (%)	23.58	23.81	18.39	0.970	0.383	0.364	0.607
Other field (%)	11.32	10.48	11.49	0.845	0.970	0.823	0.971

Note: This table presents means for baseline covariates across treatment arms. Columns report means for Private (N = 106), Public (N = 105), and Threshold (N = 87) treatments. Pairwise p-values test equality of means using two-sample t-tests. The joint p-value tests equality across all three treatments using ANOVA F-tests. Political ideology is measured on a 7-point scale from "Very Conservative" (1) to "Very Liberal" (7). Race/ethnicity categories are not mutually exclusive (participants could select multiple identities). Party identification includes participants who identify with or lean toward each party. Field of study reflects participants' primary academic major.

Appendix Table C3: Data Quality Indicators

	Private	Public	Threshold
	(1)	(2)	(3)
Duration (minutes), median	4.8	5.5	10.0
Vote comprehension			
Pass on 1st attempt (%)	62	59	62
Pass on $\leq$ 2nd attempt (%)	94	93	91
Mean attempts	1.48	1.58	1.84
Threshold comprehension			
Pass on 1st attempt (%)	—	—	79
Pass on $\leq$ 2nd attempt (%)	—	—	92
Mean attempts	—	—	1.33
Attrition rate (%)	7.8	7.1	13.0
Number of respondents	106	105	87

Note: This table shows data quality metrics for the analytic sample (N = 298). Columns (1)–(3) report statistics by treatment arm (Private, Public, Threshold). Attrition rate is the share of randomized participants who did not complete the survey; attrition did not differ significantly across treatments (Fisher's exact $p = 0.281$). Duration is from survey start to completion. Vote comprehension assesses understanding of vote revelation rules under each mechanism (all treatments). Threshold comprehension assesses understanding of threshold-based disclosure (Threshold only; columns 1–2 show "—"). Participants could retry comprehension checks until passing; the table reports first-attempt pass rates, pass rates within two attempts, and mean attempts.

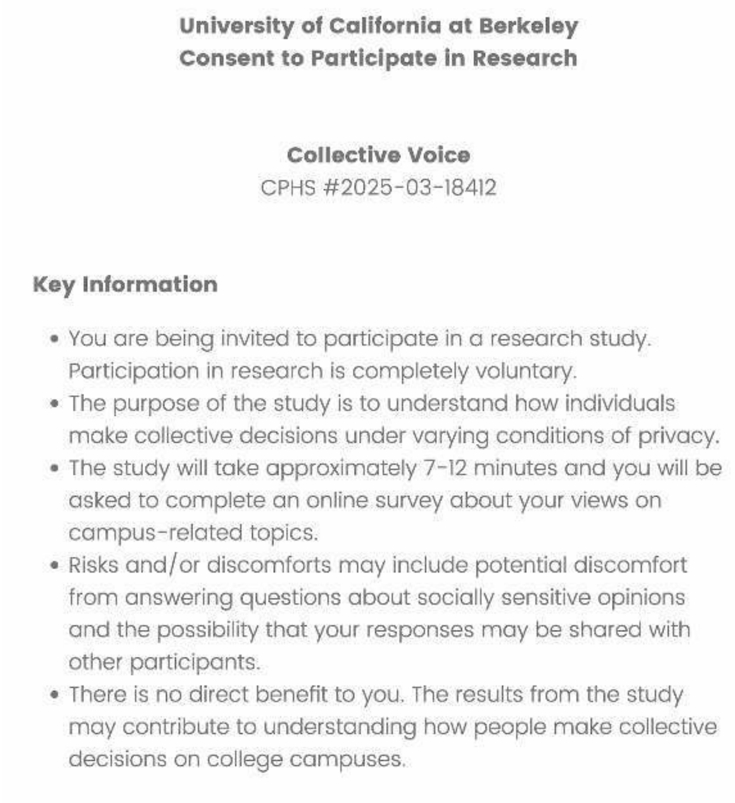
D Experimental Instructions and Survey Interface

D.1 Survey Screens

Table D1 presents screenshots of the survey interface in the order participants encountered them. Bold text above indicates branching. The experimental flow of the survey can be read off from Figure 1 in the main text.

Appendix Table D1: Survey Screens

Screening



University of California at Berkeley
Consent to Participate in Research

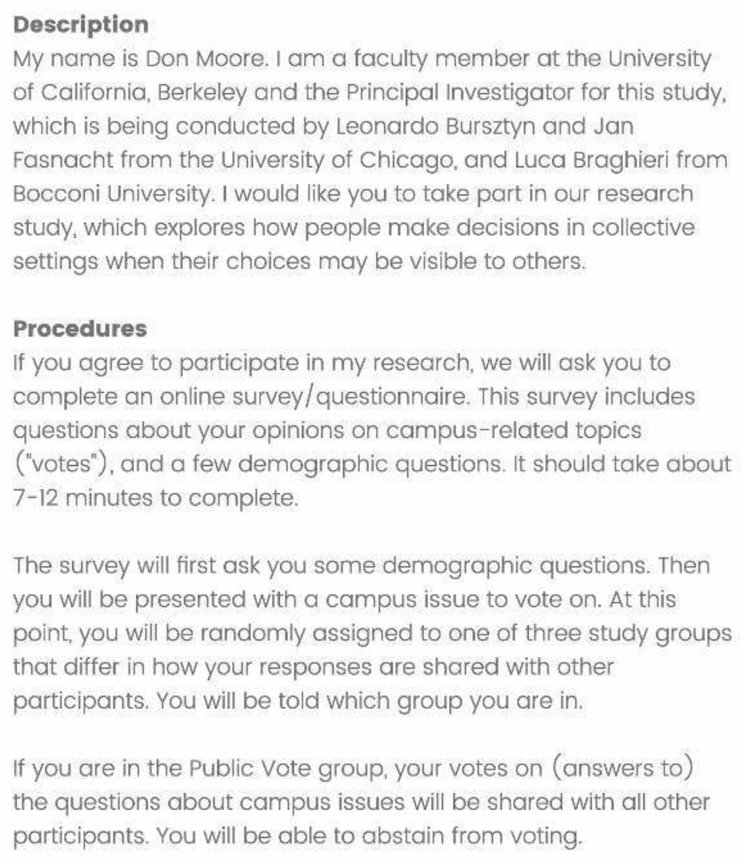
Collective Voice
CPHS #2025-03-18412

Key Information

- You are being invited to participate in a research study. Participation in research is completely voluntary.
- The purpose of the study is to understand how individuals make collective decisions under varying conditions of privacy.
- The study will take approximately 7-12 minutes and you will be asked to complete an online survey about your views on campus-related topics.
- Risks and/or discomforts may include potential discomfort from answering questions about socially sensitive opinions and the possibility that your responses may be shared with other participants.
- There is no direct benefit to you. The results from the study may contribute to understanding how people make collective decisions on college campuses.

Continued on next page

Appendix Table D1 – continued from previous page



Description

My name is Don Moore. I am a faculty member at the University of California, Berkeley and the Principal Investigator for this study, which is being conducted by Leonardo Bursztyn and Jan Fasnacht from the University of Chicago, and Luca Braghieri from Bocconi University. I would like you to take part in our research study, which explores how people make decisions in collective settings when their choices may be visible to others.

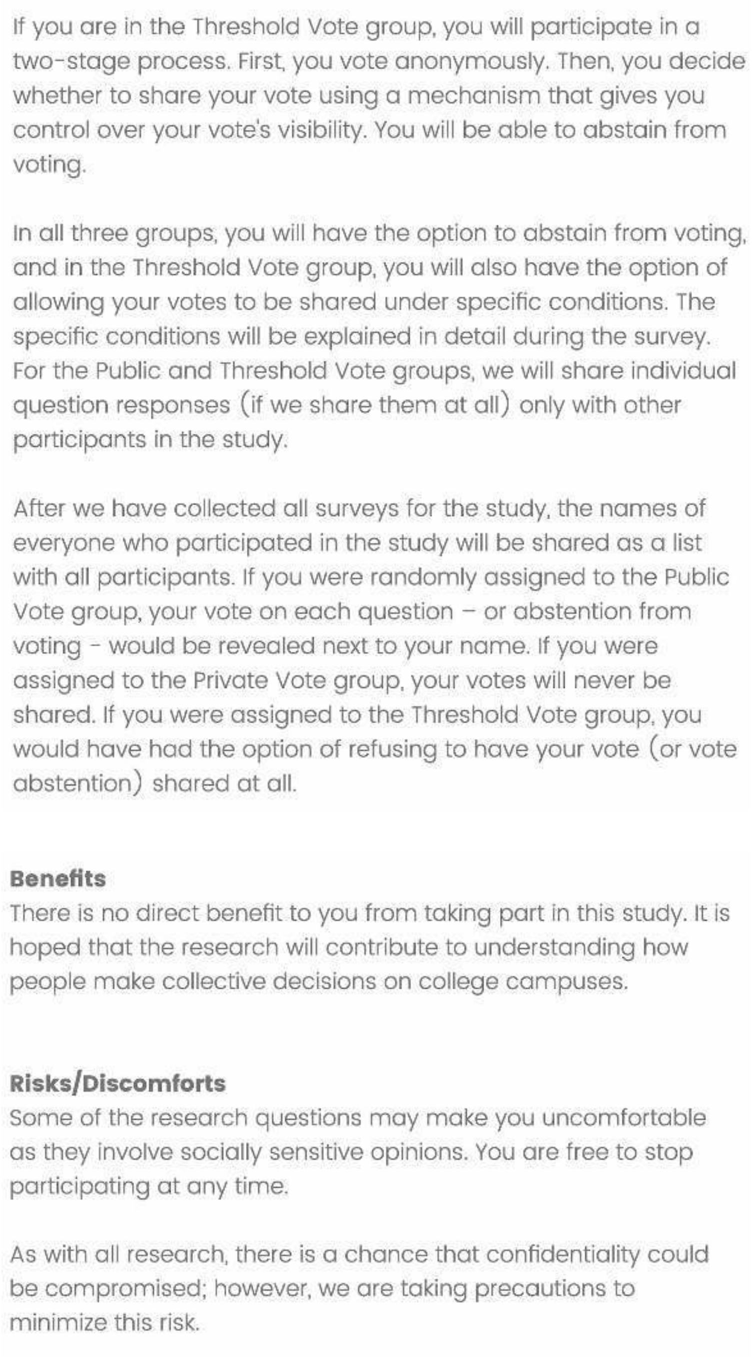
Procedures

If you agree to participate in my research, we will ask you to complete an online survey/questionnaire. This survey includes questions about your opinions on campus-related topics ("votes"), and a few demographic questions. It should take about 7-12 minutes to complete.

The survey will first ask you some demographic questions. Then you will be presented with a campus issue to vote on. At this point, you will be randomly assigned to one of three study groups that differ in how your responses are shared with other participants. You will be told which group you are in.

If you are in the Public Vote group, your votes on (answers to) the questions about campus issues will be shared with all other participants. You will be able to abstain from voting.

Continued on next page



If you are in the Threshold Vote group, you will participate in a two-stage process. First, you vote anonymously. Then, you decide whether to share your vote using a mechanism that gives you control over your vote's visibility. You will be able to abstain from voting.

In all three groups, you will have the option to abstain from voting, and in the Threshold Vote group, you will also have the option of allowing your votes to be shared under specific conditions. The specific conditions will be explained in detail during the survey. For the Public and Threshold Vote groups, we will share individual question responses (if we share them at all) only with other participants in the study.

After we have collected all surveys for the study, the names of everyone who participated in the study will be shared as a list with all participants. If you were randomly assigned to the Public Vote group, your vote on each question – or abstention from voting - would be revealed next to your name. If you were assigned to the Private Vote group, your votes will never be shared. If you were assigned to the Threshold Vote group, you would have had the option of refusing to have your vote (or vote abstention) shared at all.

Benefits

There is no direct benefit to you from taking part in this study. It is hoped that the research will contribute to understanding how people make collective decisions on college campuses.

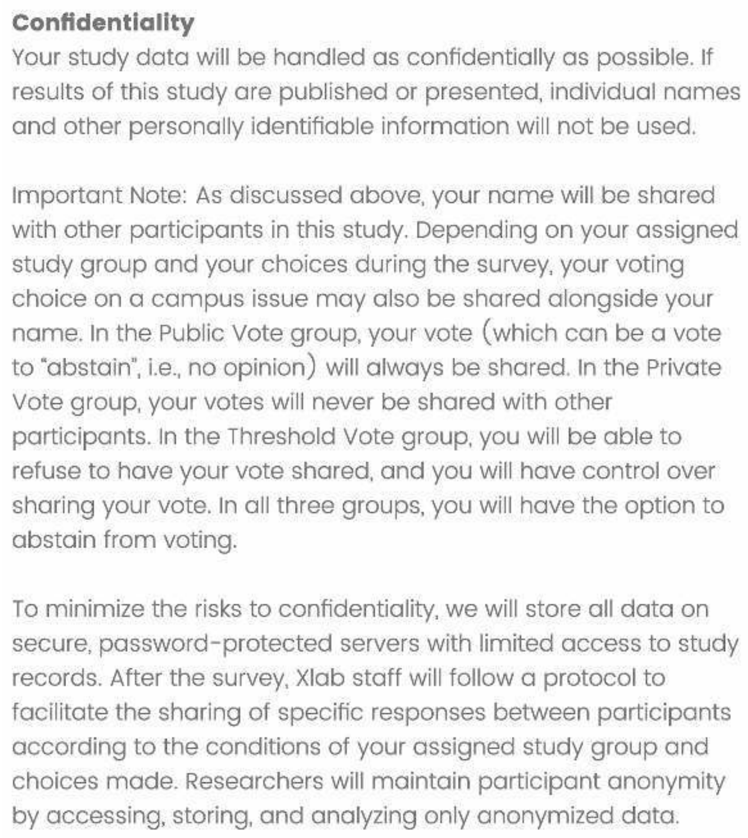
Risks/Discomforts

Some of the research questions may make you uncomfortable as they involve socially sensitive opinions. You are free to stop participating at any time.

As with all research, there is a chance that confidentiality could be compromised; however, we are taking precautions to minimize this risk.

Continued on next page

Appendix Table D1 – continued from previous page

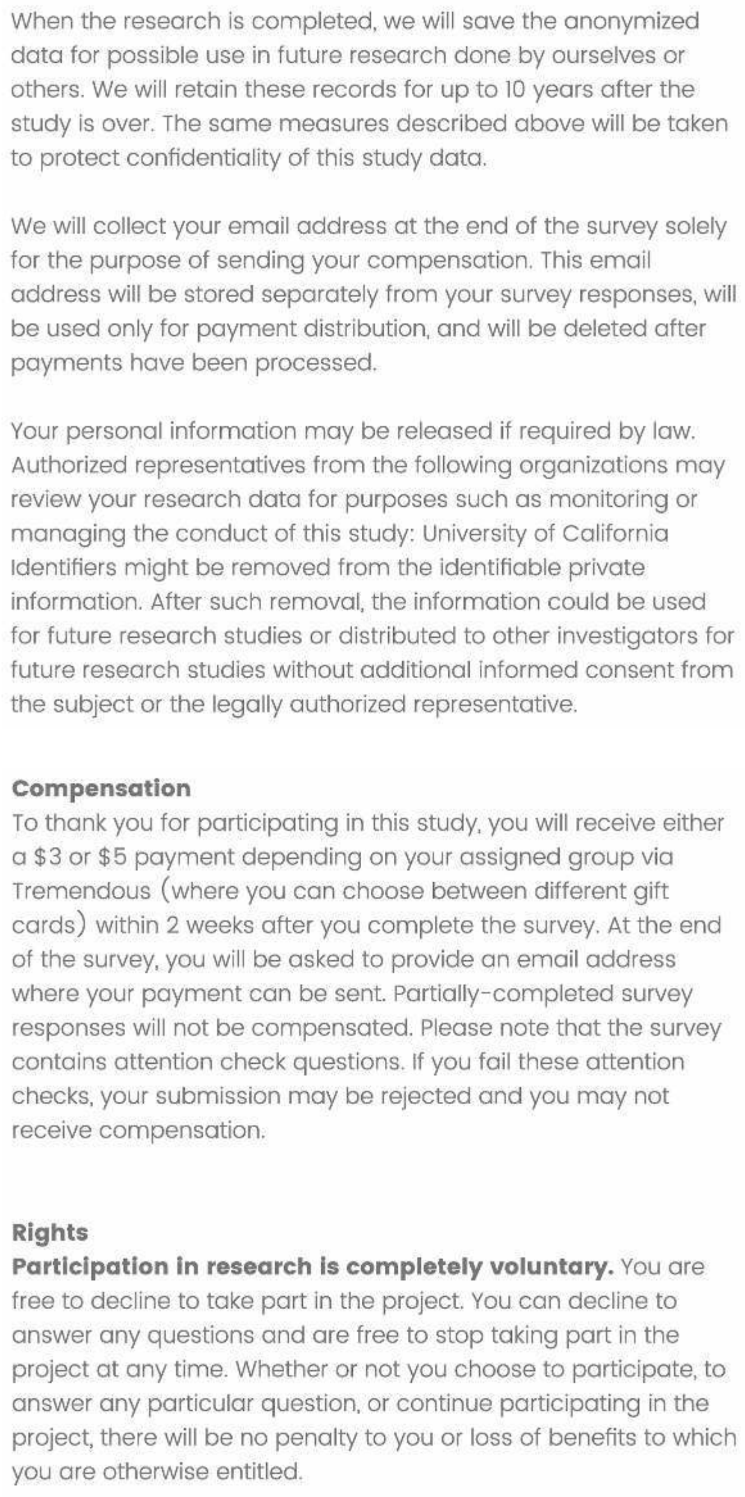


Confidentiality
Your study data will be handled as confidentially as possible. If results of this study are published or presented, individual names and other personally identifiable information will not be used.

Important Note: As discussed above, your name will be shared with other participants in this study. Depending on your assigned study group and your choices during the survey, your voting choice on a campus issue may also be shared alongside your name. In the Public Vote group, your vote (which can be a vote to "abstain", i.e., no opinion) will always be shared. In the Private Vote group, your votes will never be shared with other participants. In the Threshold Vote group, you will be able to refuse to have your vote shared, and you will have control over sharing your vote. In all three groups, you will have the option to abstain from voting.

To minimize the risks to confidentiality, we will store all data on secure, password-protected servers with limited access to study records. After the survey, Xlab staff will follow a protocol to facilitate the sharing of specific responses between participants according to the conditions of your assigned study group and choices made. Researchers will maintain participant anonymity by accessing, storing, and analyzing only anonymized data.

Continued on next page



When the research is completed, we will save the anonymized data for possible use in future research done by ourselves or others. We will retain these records for up to 10 years after the study is over. The same measures described above will be taken to protect confidentiality of this study data.

We will collect your email address at the end of the survey solely for the purpose of sending your compensation. This email address will be stored separately from your survey responses, will be used only for payment distribution, and will be deleted after payments have been processed.

Your personal information may be released if required by law. Authorized representatives from the following organizations may review your research data for purposes such as monitoring or managing the conduct of this study: University of California Identifiers might be removed from the identifiable private information. After such removal, the information could be used for future research studies or distributed to other investigators for future research studies without additional informed consent from the subject or the legally authorized representative.

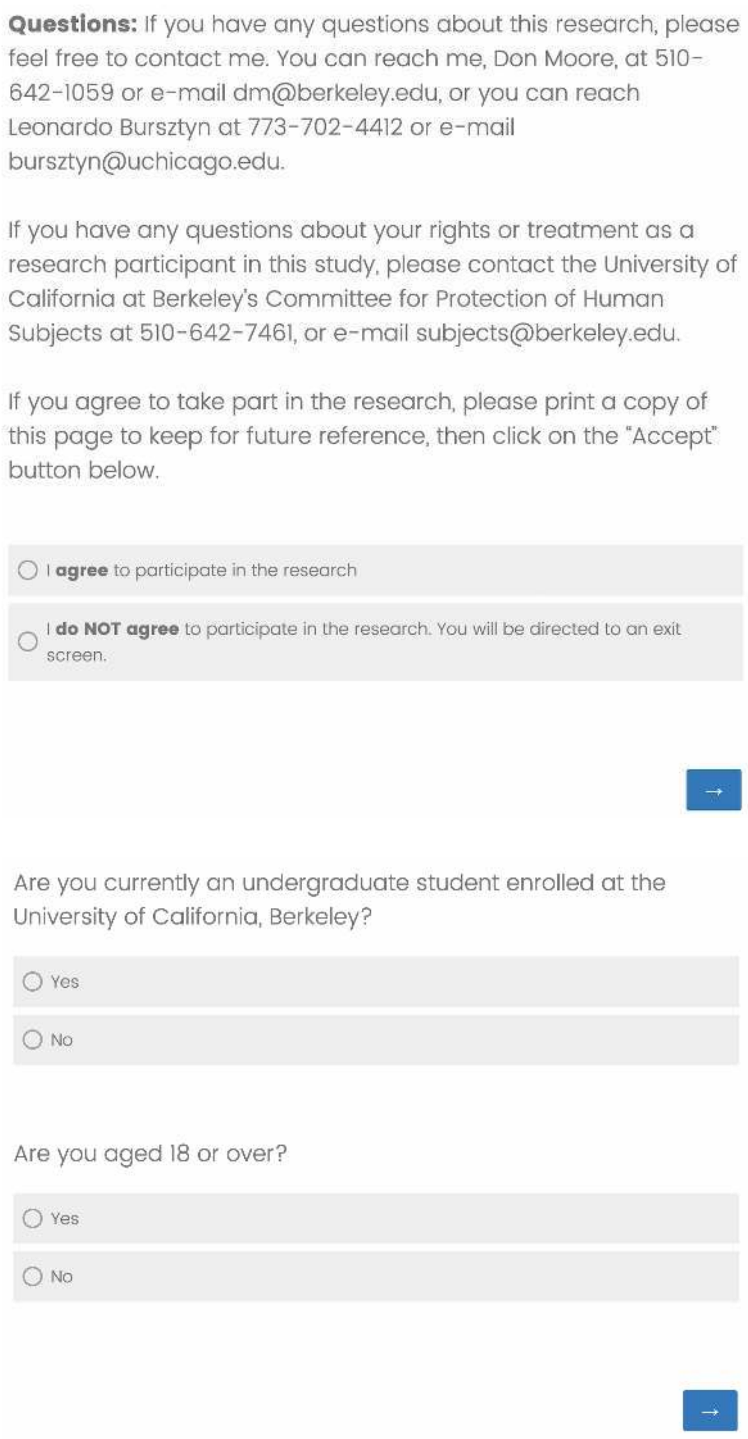
Compensation

To thank you for participating in this study, you will receive either a $3 or $5 payment depending on your assigned group via Tremendous (where you can choose between different gift cards) within 2 weeks after you complete the survey. At the end of the survey, you will be asked to provide an email address where your payment can be sent. Partially-completed survey responses will not be compensated. Please note that the survey contains attention check questions. If you fail these attention checks, your submission may be rejected and you may not receive compensation.

Rights

Participation in research is completely voluntary. You are free to decline to take part in the project. You can decline to answer any questions and are free to stop taking part in the project at any time. Whether or not you choose to participate, to answer any particular question, or continue participating in the project, there will be no penalty to you or loss of benefits to which you are otherwise entitled.

Continued on next page



Questions: If you have any questions about this research, please feel free to contact me. You can reach me, Don Moore, at 510-642-1059 or e-mail dm@berkeley.edu, or you can reach Leonardo Bursztyn at 773-702-4412 or e-mail bursztyn@uchicago.edu.

If you have any questions about your rights or treatment as a research participant in this study, please contact the University of California at Berkeley's Committee for Protection of Human Subjects at 510-642-7461, or e-mail subjects@berkeley.edu.

If you agree to take part in the research, please print a copy of this page to keep for future reference, then click on the "Accept" button below.

○ I **agree** to participate in the research

○ I **do NOT agree** to participate in the research. You will be directed to an exit screen.

Are you currently an undergraduate student enrolled at the University of California, Berkeley?

○ Yes

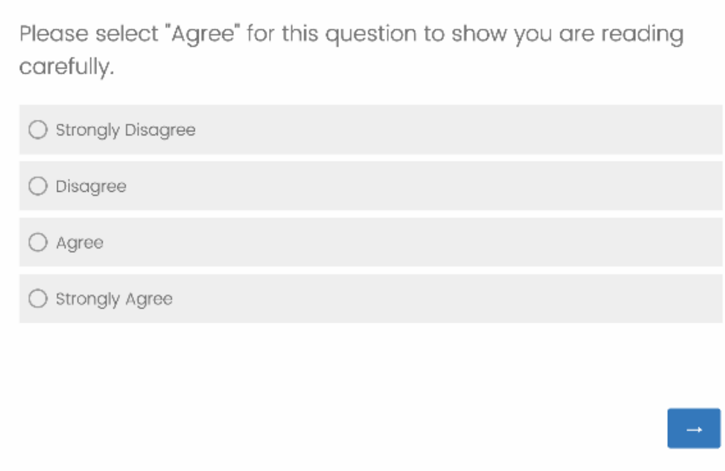
○ No

Are you aged 18 or over?

○ Yes

○ No

Continued on next page



Please select "Agree" for this question to show you are reading carefully.

○ Strongly Disagree

○ Disagree

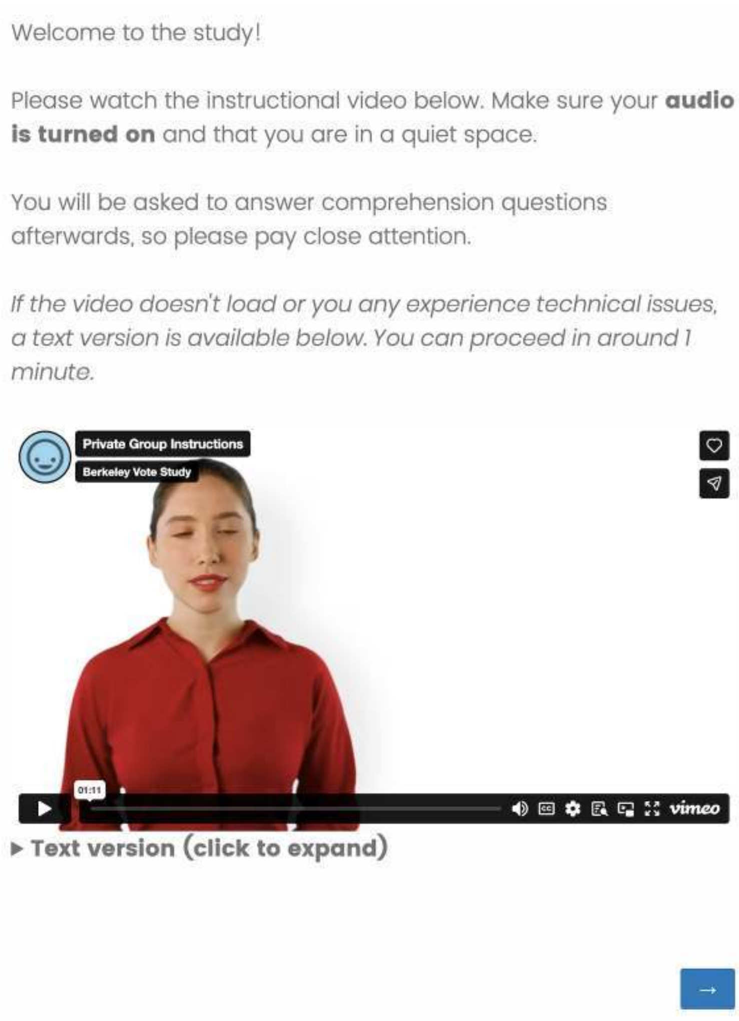
○ Agree

○ Strongly Agree

Instructional video (treatment-specific)

Private treatment:

Continued on next page



Welcome to the study!

Please watch the instructional video below. Make sure your **audio is turned on** and that you are in a quiet space.

You will be asked to answer comprehension questions afterwards, so please pay close attention.

If the video doesn't load or you any experience technical issues, a text version is available below. You can proceed in around 1 minute.

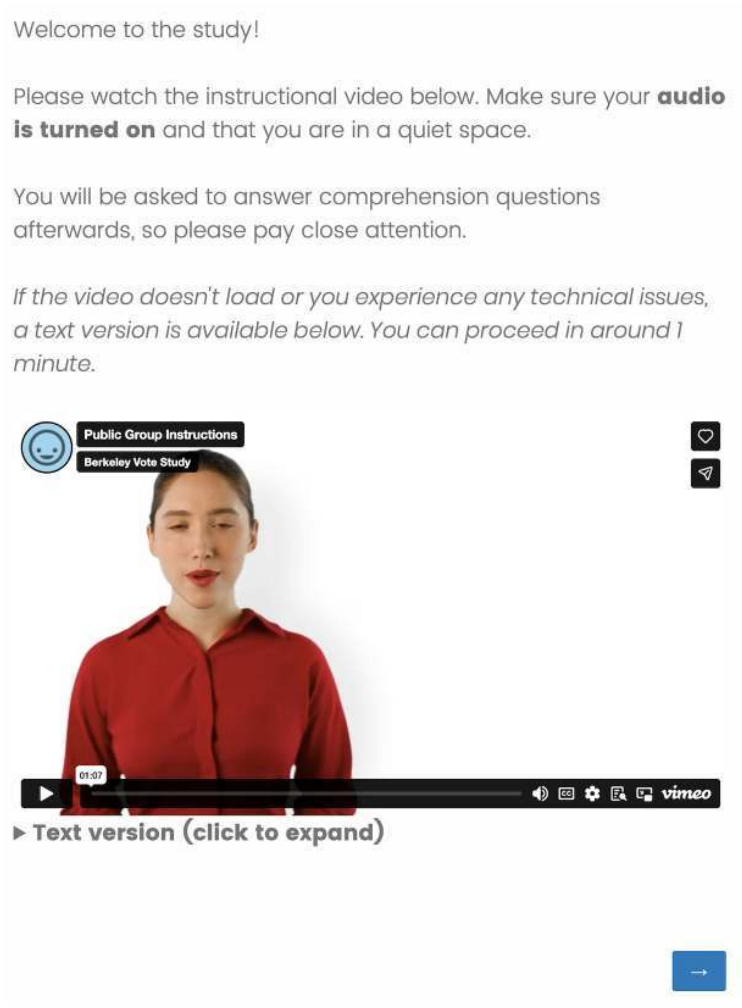
▸ **Text version (click to expand)**

→

Public treatment:

Continued on next page

Appendix Table D1 – continued from previous page



Welcome to the study!

Please watch the instructional video below. Make sure your **audio is turned on** and that you are in a quiet space.

You will be asked to answer comprehension questions afterwards, so please pay close attention.

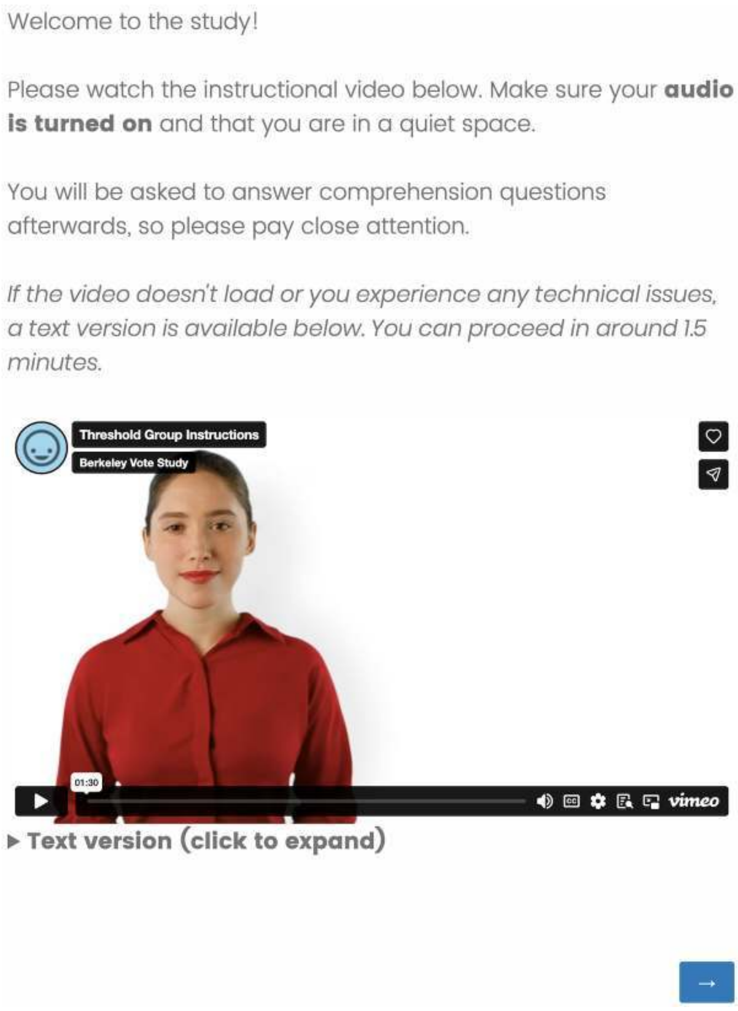
If the video doesn't load or you experience any technical issues, a text version is available below. You can proceed in around 1 minute.

▶ **Text version (click to expand)**

Threshold treatment:

Continued on next page

Appendix Table D1 – continued from previous page



Welcome to the study!

Please watch the instructional video below. Make sure your **audio is turned on** and that you are in a quiet space.

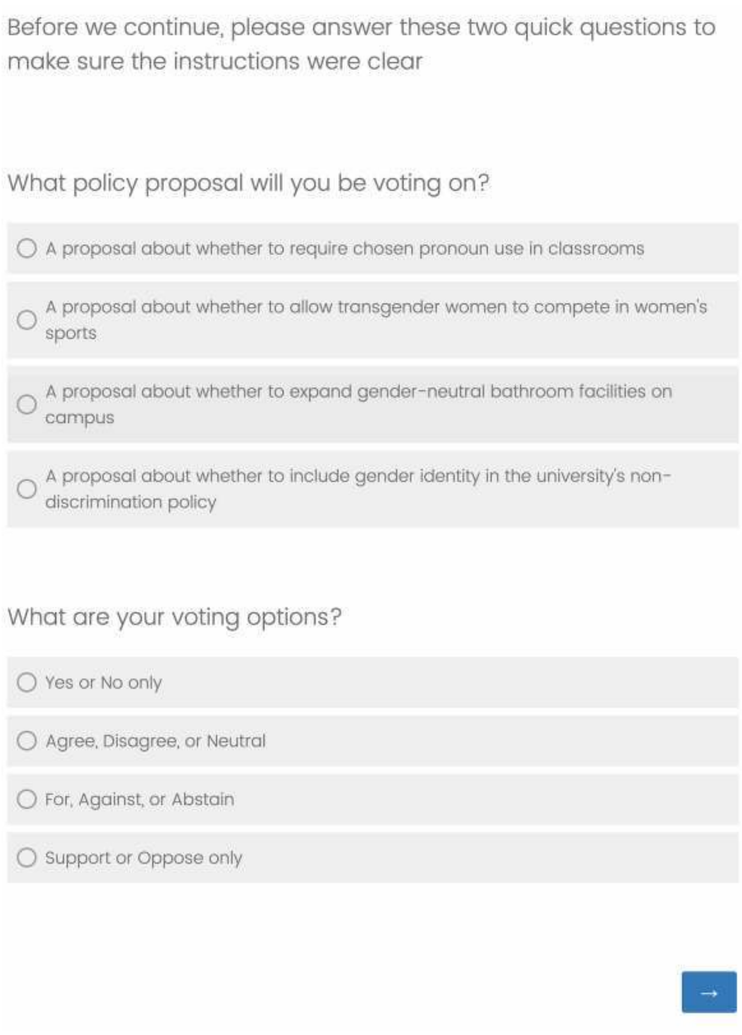
You will be asked to answer comprehension questions afterwards, so please pay close attention.

If the video doesn't load or you experience any technical issues, a text version is available below. You can proceed in around 1.5 minutes.

▸ **Text version (click to expand)**

Video comprehension check

Continued on next page



Before we continue, please answer these two quick questions to make sure the instructions were clear

What policy proposal will you be voting on?

○ A proposal about whether to require chosen pronoun use in classrooms

○ A proposal about whether to allow transgender women to compete in women's sports

○ A proposal about whether to expand gender-neutral bathroom facilities on campus

○ A proposal about whether to include gender identity in the university's non-discrimination policy

What are your voting options?

○ Yes or No only

○ Agree, Disagree, or Neutral

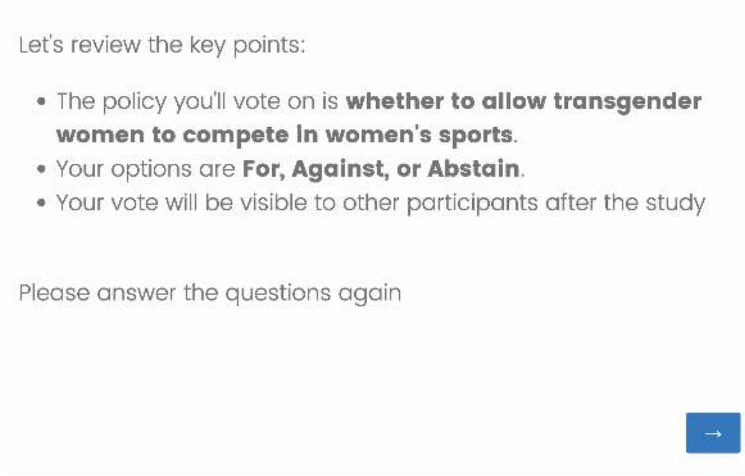
○ For, Against, or Abstain

○ Support or Oppose only

If incorrect:

Continued on next page

Appendix Table D1 – continued from previous page



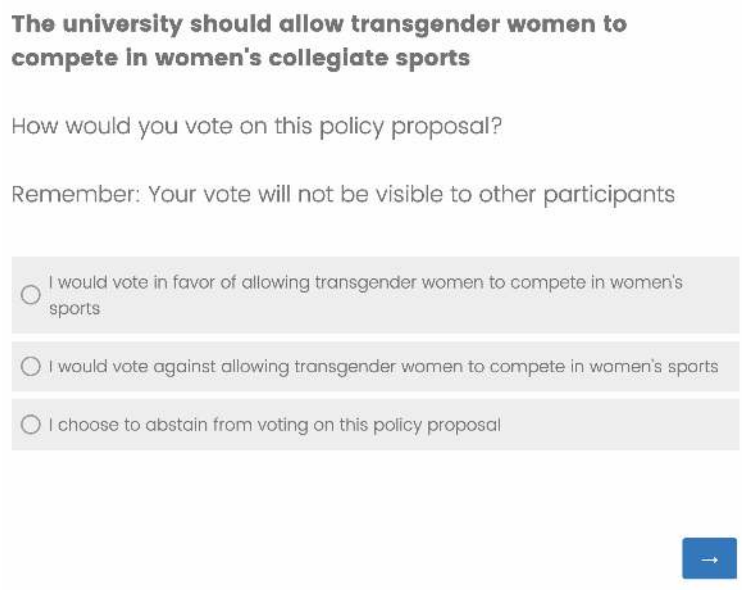
Let's review the key points:

- The policy you'll vote on is **whether to allow transgender women to compete in women's sports**.
- Your options are **For, Against, or Abstain**.
- Your vote will be visible to other participants after the study

Please answer the questions again

Vote decision

Private treatment:



The university should allow transgender women to compete in women's collegiate sports

How would you vote on this policy proposal?

Remember: Your vote will not be visible to other participants

○ I would vote in favor of allowing transgender women to compete in women's sports

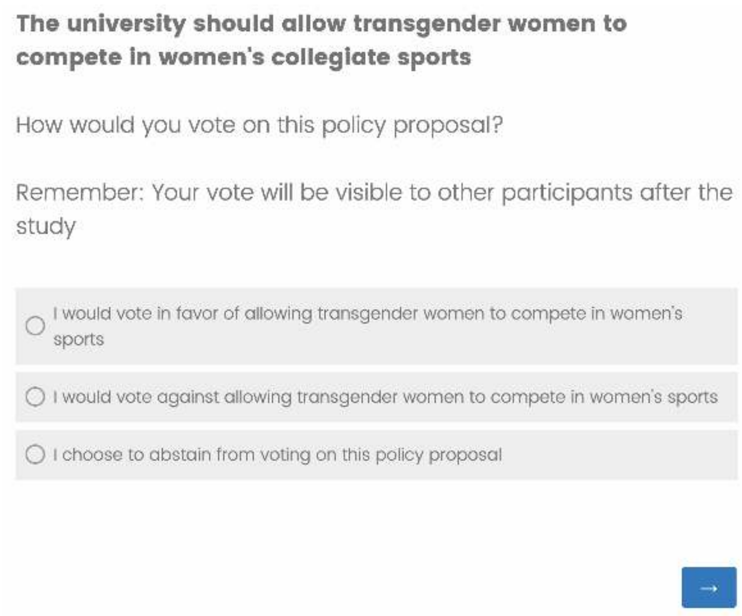
○ I would vote against allowing transgender women to compete in women's sports

○ I choose to abstain from voting on this policy proposal

Public treatment:

Continued on next page

Appendix Table D1 – continued from previous page



The university should allow transgender women to compete in women's collegiate sports

How would you vote on this policy proposal?

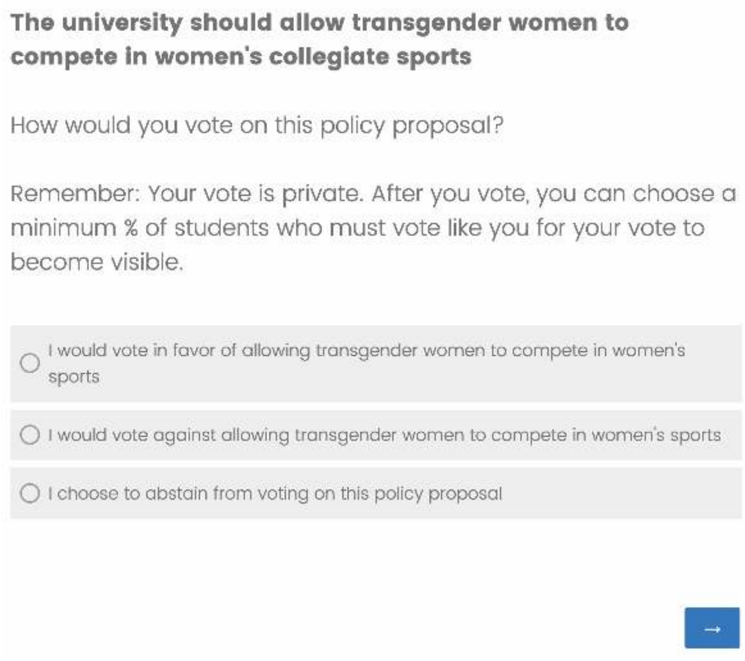
Remember: Your vote will be visible to other participants after the study

○ I would vote in favor of allowing transgender women to compete in women's sports

○ I would vote against allowing transgender women to compete in women's sports

○ I choose to abstain from voting on this policy proposal

Threshold treatment:



The university should allow transgender women to compete in women's collegiate sports

How would you vote on this policy proposal?

Remember: Your vote is private. After you vote, you can choose a minimum % of students who must vote like you for your vote to become visible.

○ I would vote in favor of allowing transgender women to compete in women's sports

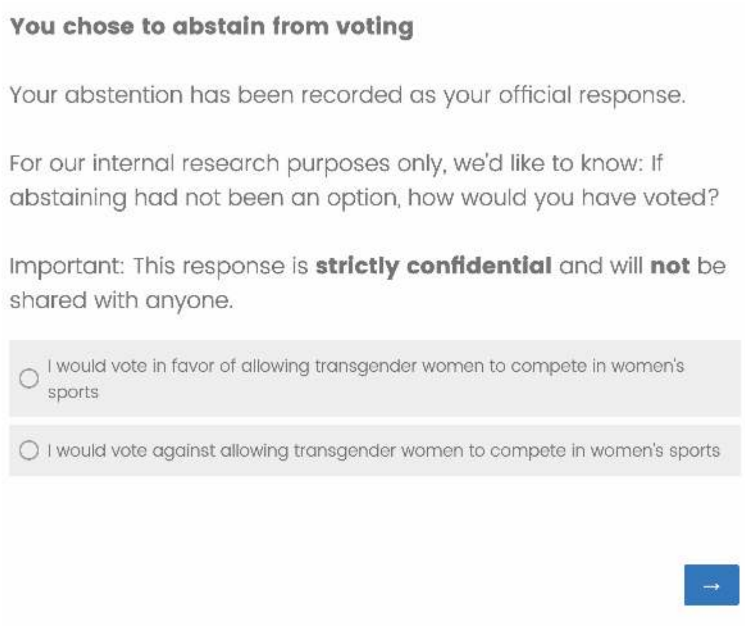
○ I would vote against allowing transgender women to compete in women's sports

○ I choose to abstain from voting on this policy proposal

If abstained:

Continued on next page

Appendix Table D1 – continued from previous page



You chose to abstain from voting

Your abstention has been recorded as your official response.

For our internal research purposes only, we'd like to know: If abstaining had not been an option, how would you have voted?

Important: This response is **strictly confidential** and will **not** be shared with anyone.

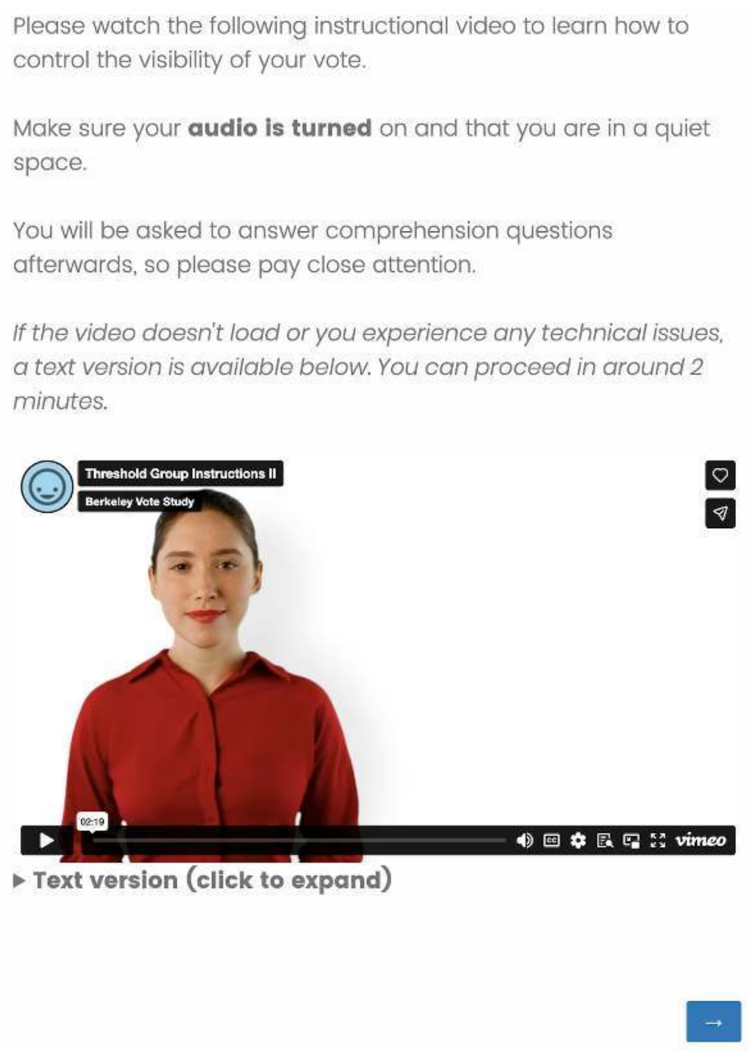
○ I would vote in favor of allowing transgender women to compete in women's sports

○ I would vote against allowing transgender women to compete in women's sports

Threshold task (Non-abstaining Threshold treatment participants only)

Continued on next page

Appendix Table D1 – continued from previous page



Please watch the following instructional video to learn how to control the visibility of your vote.

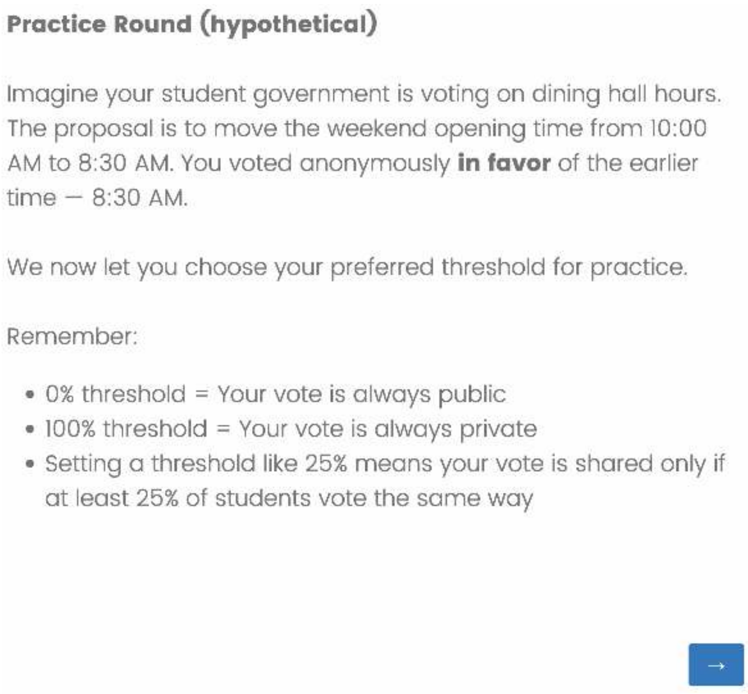
Make sure your **audio is turned** on and that you are in a quiet space.

You will be asked to answer comprehension questions afterwards, so please pay close attention.

If the video doesn't load or you experience any technical issues, a text version is available below. You can proceed in around 2 minutes.

▸ **Text version (click to expand)**

Continued on next page



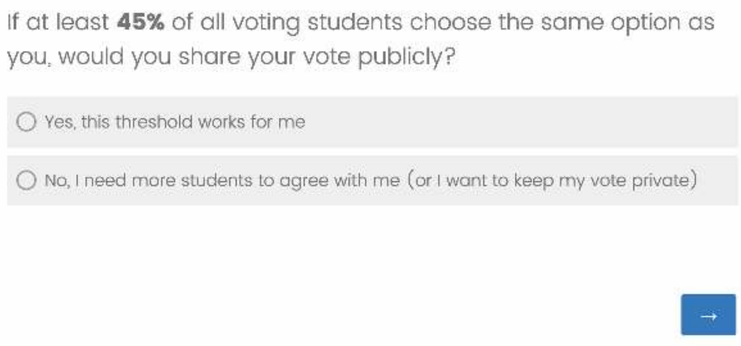
Practice Round (hypothetical)

Imagine your student government is voting on dining hall hours. The proposal is to move the weekend opening time from 10:00 AM to 8:30 AM. You voted anonymously **in favor** of the earlier time — 8:30 AM.

We now let you choose your preferred threshold for practice.

Remember:

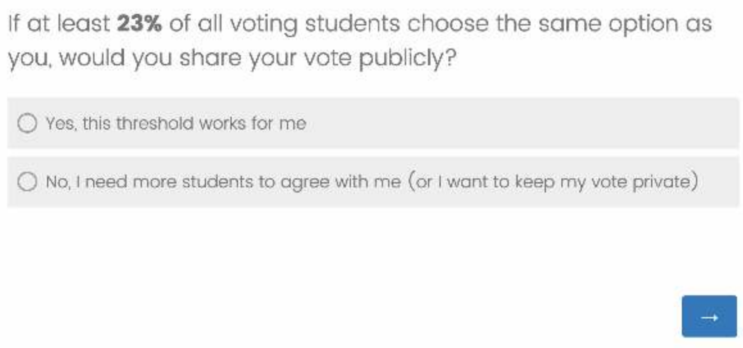
- 0% threshold = Your vote is always public
- 100% threshold = Your vote is always private
- Setting a threshold like 25% means your vote is shared only if at least 25% of students vote the same way



If at least **45%** of all voting students choose the same option as you, would you share your vote publicly?

○ Yes, this threshold works for me

○ No, I need more students to agree with me (or I want to keep my vote private)



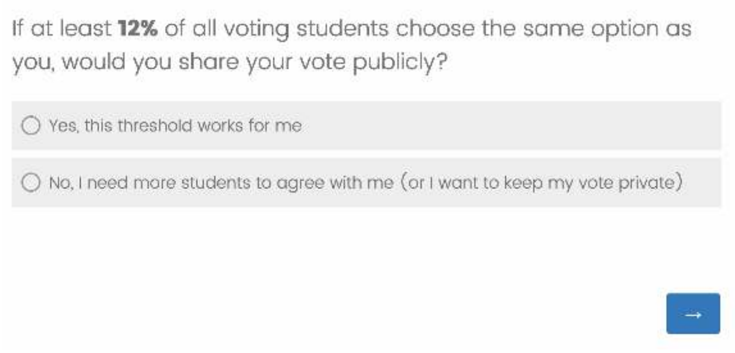
If at least **23%** of all voting students choose the same option as you, would you share your vote publicly?

○ Yes, this threshold works for me

○ No, I need more students to agree with me (or I want to keep my vote private)

Continued on next page

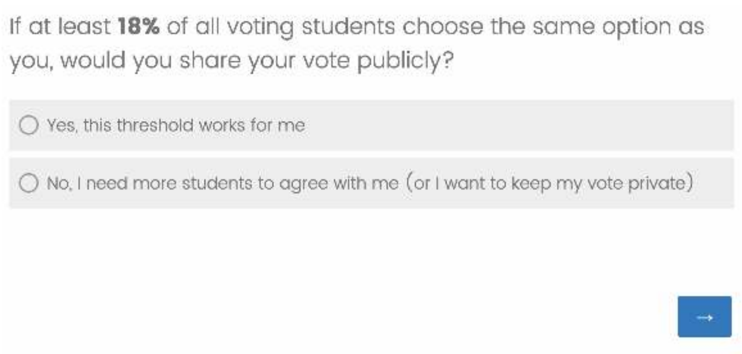
Appendix Table D1 – continued from previous page



If at least **12%** of all voting students choose the same option as you, would you share your vote publicly?

○ Yes, this threshold works for me

○ No, I need more students to agree with me (or I want to keep my vote private)



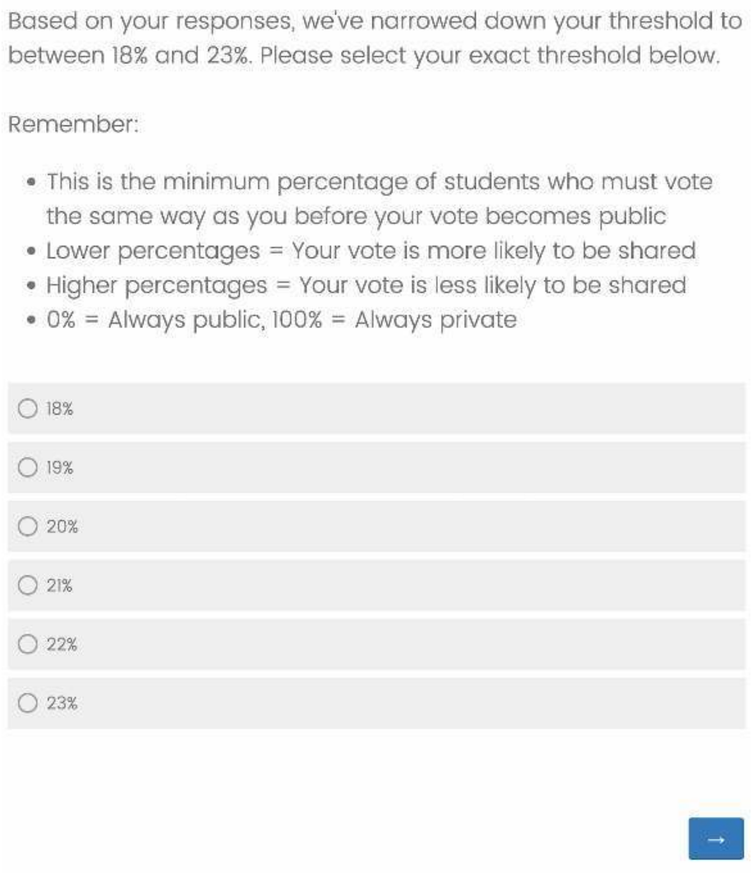
If at least **18%** of all voting students choose the same option as you, would you share your vote publicly?

○ Yes, this threshold works for me

○ No, I need more students to agree with me (or I want to keep my vote private)

Continued on next page

Appendix Table D1 – continued from previous page



Based on your responses, we've narrowed down your threshold to between 18% and 23%. Please select your exact threshold below.

Remember:

- This is the minimum percentage of students who must vote the same way as you before your vote becomes public
- Lower percentages = Your vote is more likely to be shared
- Higher percentages = Your vote is less likely to be shared
- 0% = Always public, 100% = Always private

○ 18%

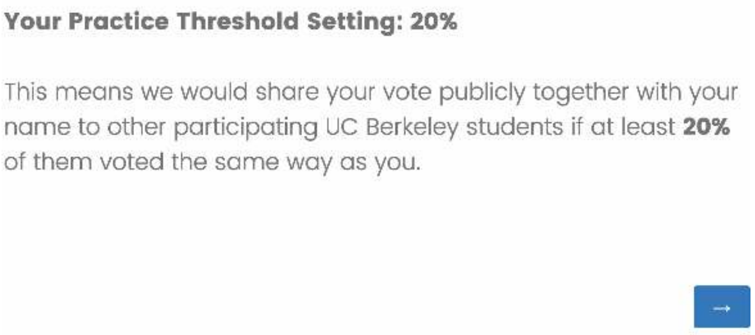
○ 19%

○ 20%

○ 21%

○ 22%

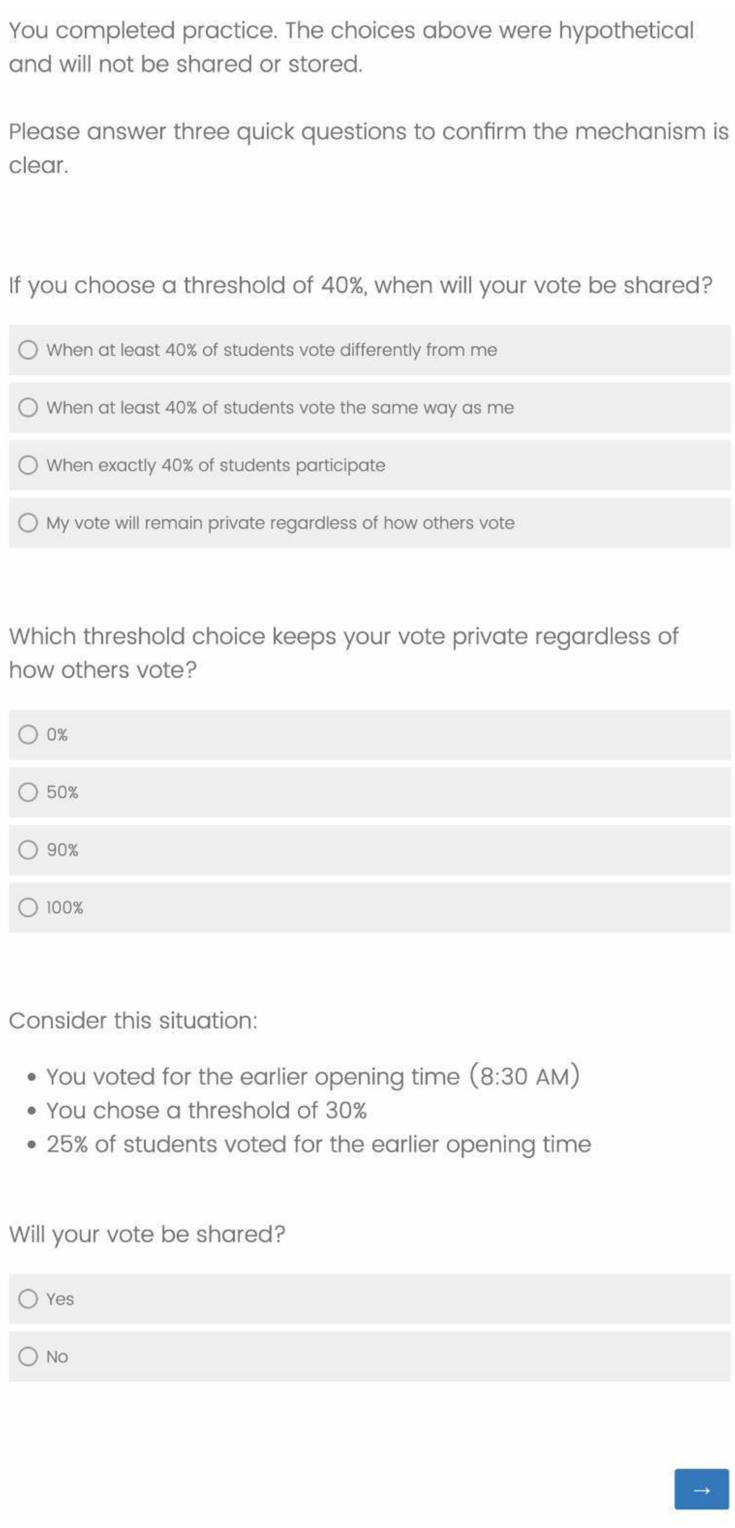
○ 23%



Your Practice Threshold Setting: 20%

This means we would share your vote publicly together with your name to other participating UC Berkeley students if at least **20%** of them voted the same way as you.

Continued on next page



You completed practice. The choices above were hypothetical and will not be shared or stored.

Please answer three quick questions to confirm the mechanism is clear.

If you choose a threshold of 40%, when will your vote be shared?

○ When at least 40% of students vote differently from me

○ When at least 40% of students vote the same way as me

○ When exactly 40% of students participate

○ My vote will remain private regardless of how others vote

Which threshold choice keeps your vote private regardless of how others vote?

○ 0%

○ 50%

○ 90%

○ 100%

Consider this situation:

- You voted for the earlier opening time (8:30 AM)
- You chose a threshold of 30%
- 25% of students voted for the earlier opening time

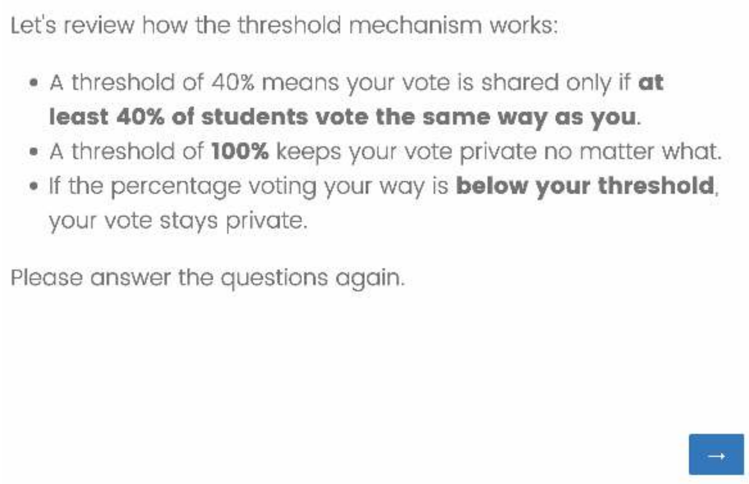
Will your vote be shared?

○ Yes

○ No

Continued on next page

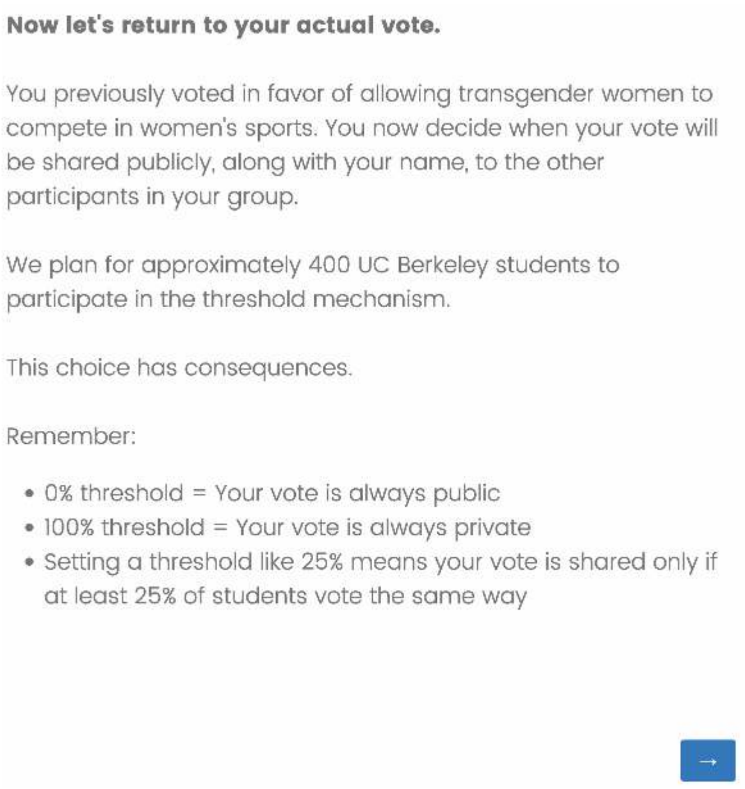
If incorrect:



Let's review how the threshold mechanism works:

- A threshold of 40% means your vote is shared only if **at least 40% of students vote the same way as you**.
- A threshold of **100%** keeps your vote private no matter what.
- If the percentage voting your way is **below your threshold**, your vote stays private.

Please answer the questions again.



Now let's return to your actual vote.

You previously voted in favor of allowing transgender women to compete in women's sports. You now decide when your vote will be shared publicly, along with your name, to the other participants in your group.

We plan for approximately 400 UC Berkeley students to participate in the threshold mechanism.

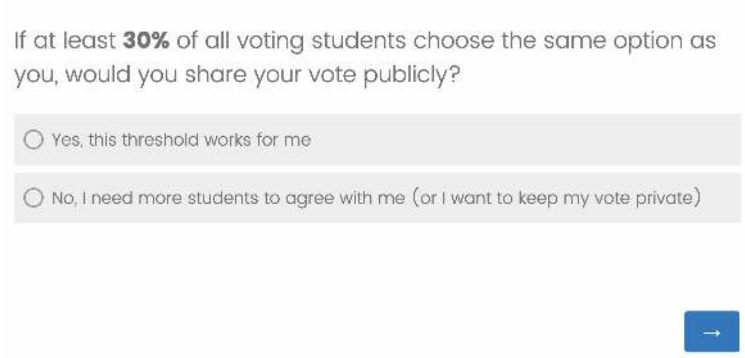
This choice has consequences.

Remember:

- 0% threshold = Your vote is always public
- 100% threshold = Your vote is always private
- Setting a threshold like 25% means your vote is shared only if at least 25% of students vote the same way

Continued on next page

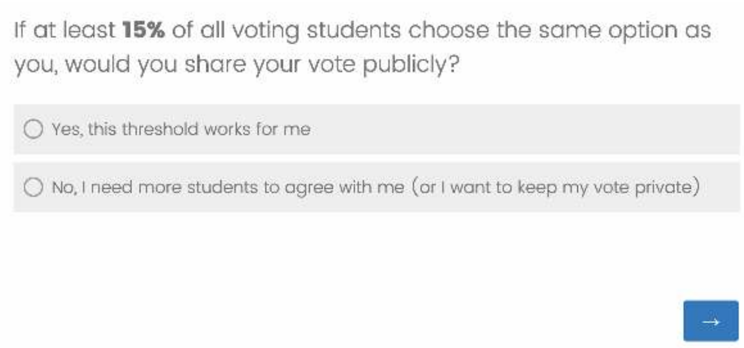
Appendix Table D1 – continued from previous page



If at least **30%** of all voting students choose the same option as you, would you share your vote publicly?

○ Yes, this threshold works for me

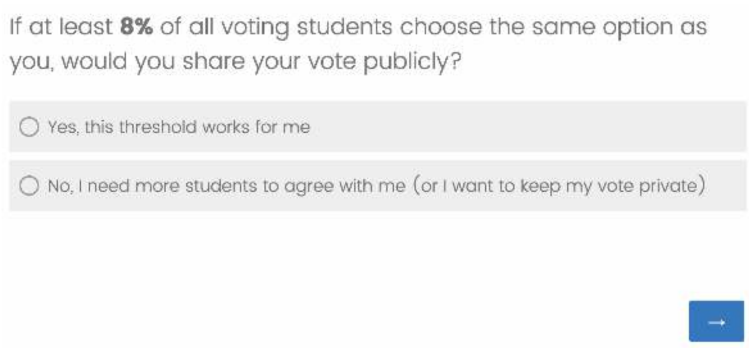
○ No, I need more students to agree with me (or I want to keep my vote private)



If at least **15%** of all voting students choose the same option as you, would you share your vote publicly?

○ Yes, this threshold works for me

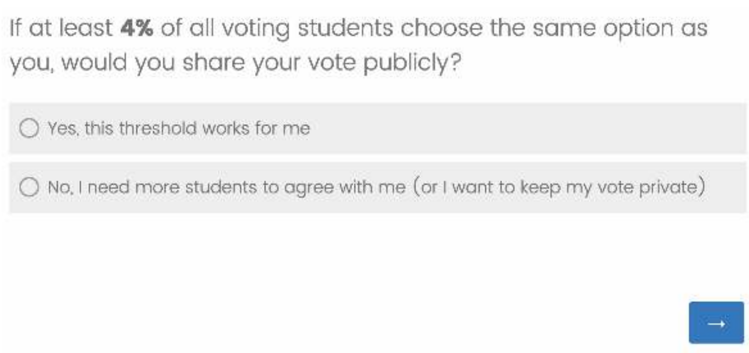
○ No, I need more students to agree with me (or I want to keep my vote private)



If at least **8%** of all voting students choose the same option as you, would you share your vote publicly?

○ Yes, this threshold works for me

○ No, I need more students to agree with me (or I want to keep my vote private)

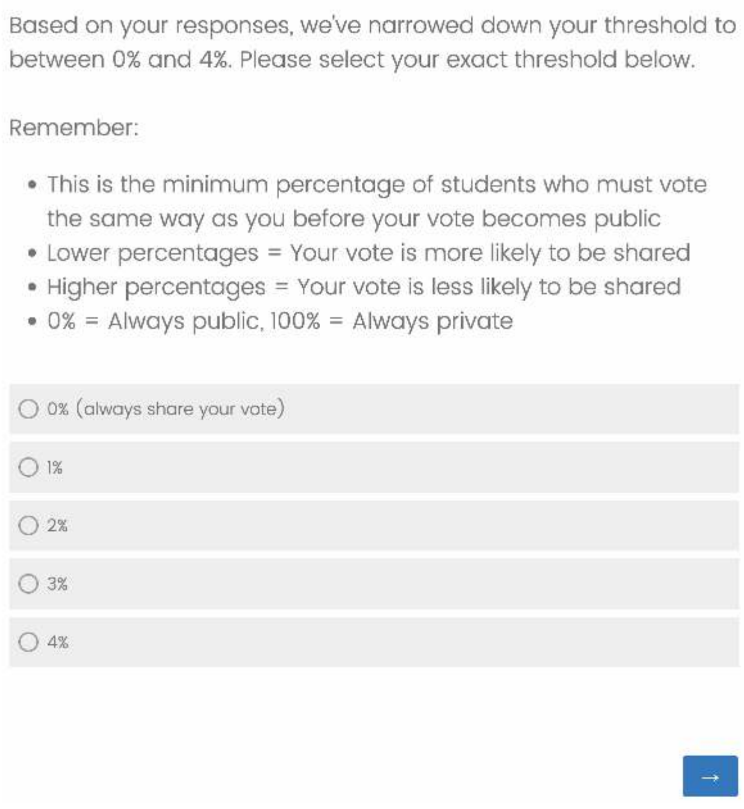


If at least **4%** of all voting students choose the same option as you, would you share your vote publicly?

○ Yes, this threshold works for me

○ No, I need more students to agree with me (or I want to keep my vote private)

Continued on next page



Based on your responses, we've narrowed down your threshold to between 0% and 4%. Please select your exact threshold below.

Remember:

- This is the minimum percentage of students who must vote the same way as you before your vote becomes public
- Lower percentages = Your vote is more likely to be shared
- Higher percentages = Your vote is less likely to be shared
- 0% = Always public, 100% = Always private

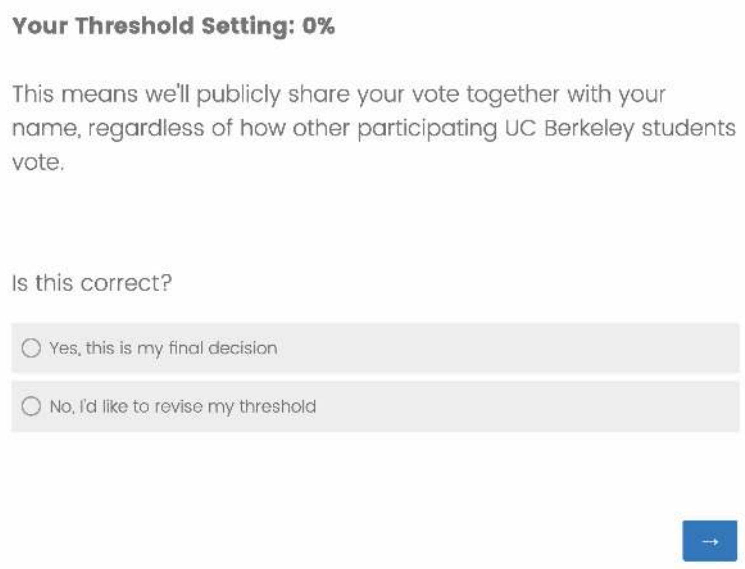
○ 0% (always share your vote)

○ 1%

○ 2%

○ 3%

○ 4%



Your Threshold Setting: 0%

This means we'll publicly share your vote together with your name, regardless of how other participating UC Berkeley students vote.

Is this correct?

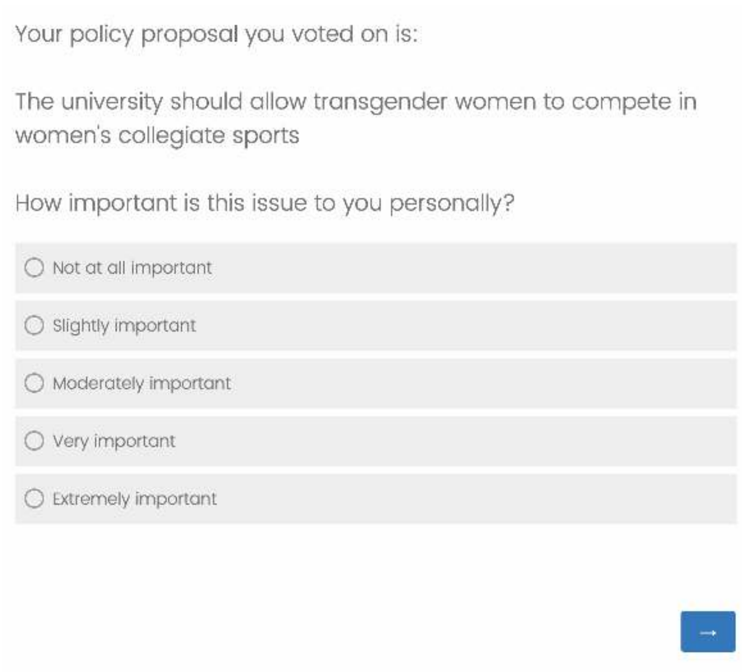
○ Yes, this is my final decision

○ No, I'd like to revise my threshold

Post-vote survey questions

Continued on next page

Appendix Table D1 – continued from previous page



Your policy proposal you voted on is:

The university should allow transgender women to compete in women's collegiate sports

How important is this issue to you personally?

○ Not at all important

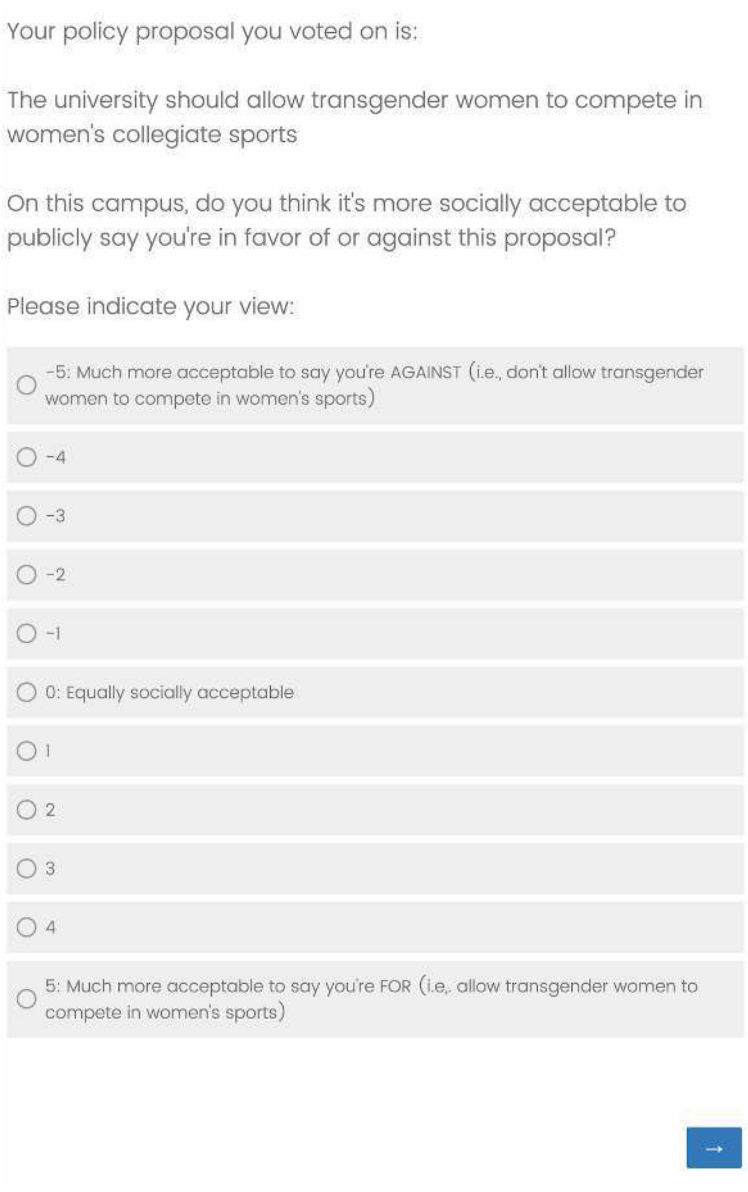
○ Slightly important

○ Moderately important

○ Very important

○ Extremely important

Continued on next page



Your policy proposal you voted on is:

The university should allow transgender women to compete in women's collegiate sports

On this campus, do you think it's more socially acceptable to publicly say you're in favor of or against this proposal?

Please indicate your view:

○ -5: Much more acceptable to say you're AGAINST (i.e., don't allow transgender women to compete in women's sports)

○ -4

○ -3

○ -2

○ -1

○ 0: Equally socially acceptable

○ 1

○ 2

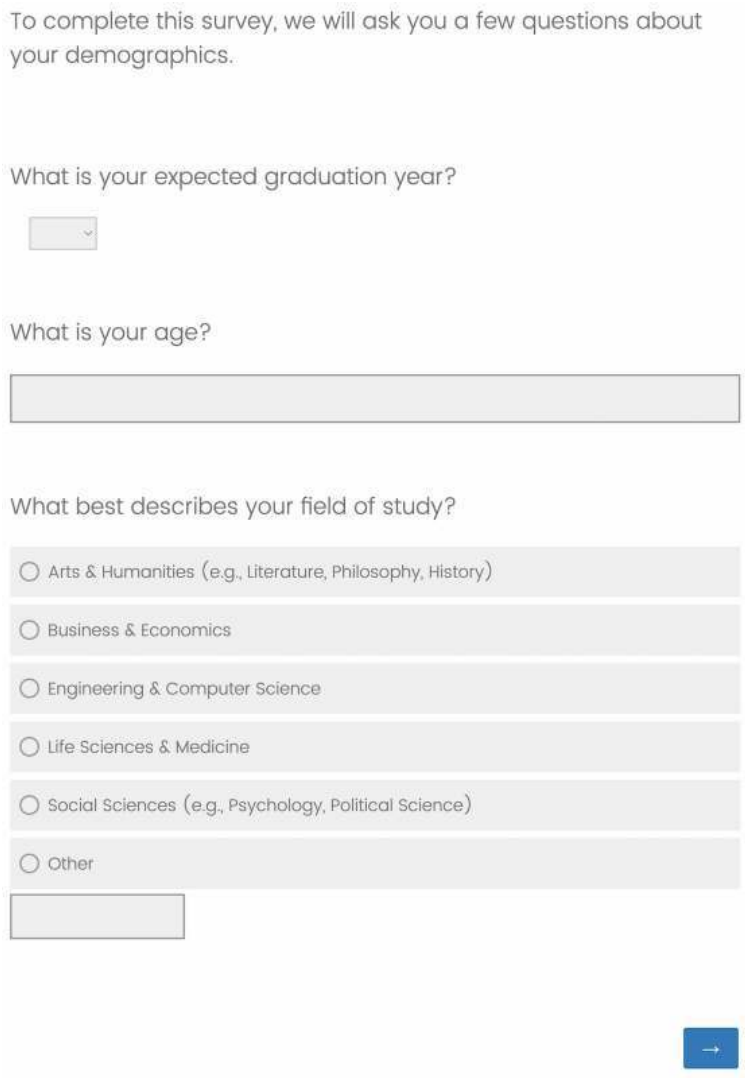
○ 3

○ 4

○ 5: Much more acceptable to say you're FOR (i.e,. allow transgender women to compete in women's sports)

Continued on next page

Appendix Table D1 – continued from previous page



To complete this survey, we will ask you a few questions about your demographics.

What is your expected graduation year?

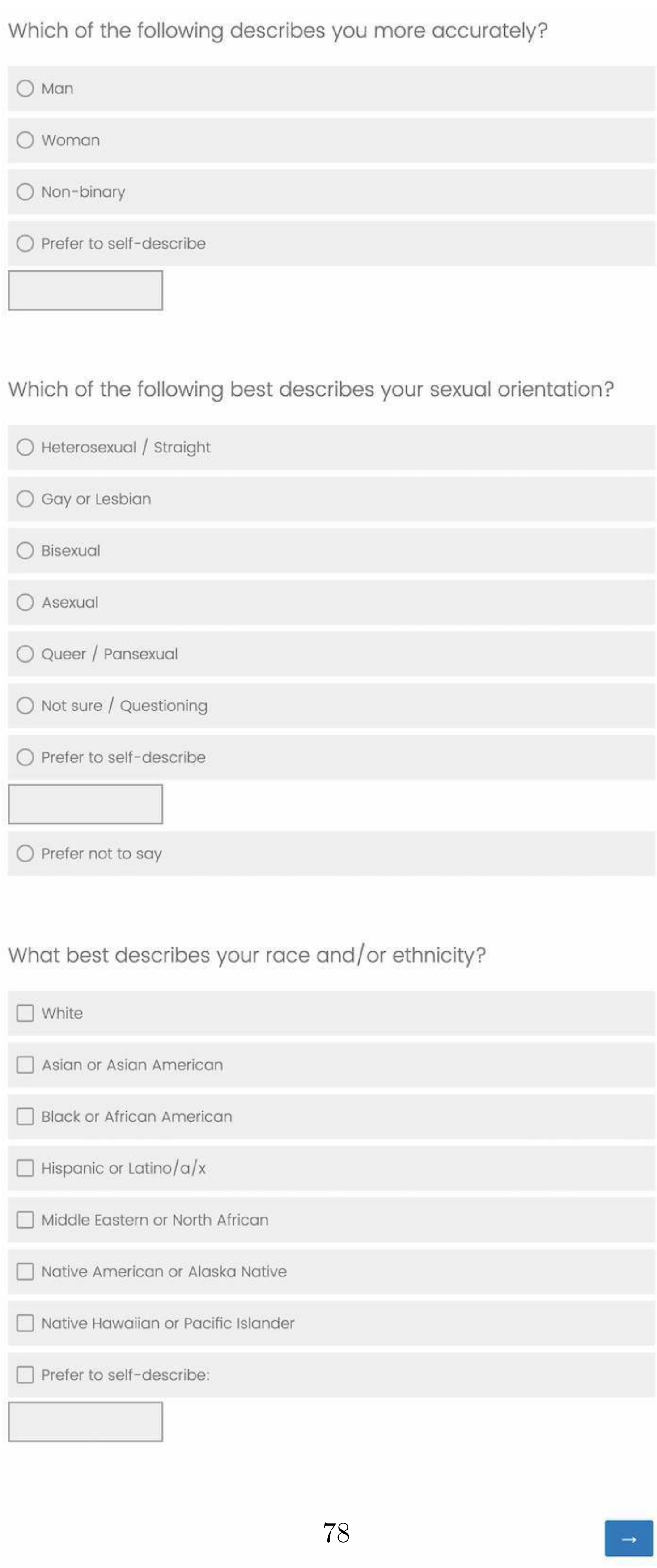
What is your age?

What best describes your field of study?

- ○ Arts & Humanities (e.g., Literature, Philosophy, History)
- ○ Business & Economics
- ○ Engineering & Computer Science
- ○ Life Sciences & Medicine
- ○ Social Sciences (e.g., Psychology, Political Science)
- ○ Other

Continued on next page

Appendix Table D1 – continued from previous page



Which of the following describes you more accurately?

○ Man

○ Woman

○ Non-binary

○ Prefer to self-describe

Which of the following best describes your sexual orientation?

○ Heterosexual / Straight

○ Gay or Lesbian

○ Bisexual

○ Asexual

○ Queer / Pansexual

○ Not sure / Questioning

○ Prefer to self-describe

○ Prefer not to say

What best describes your race and/or ethnicity?

□ White

□ Asian or Asian American

□ Black or African American

□ Hispanic or Latino/a/x

□ Middle Eastern or North African

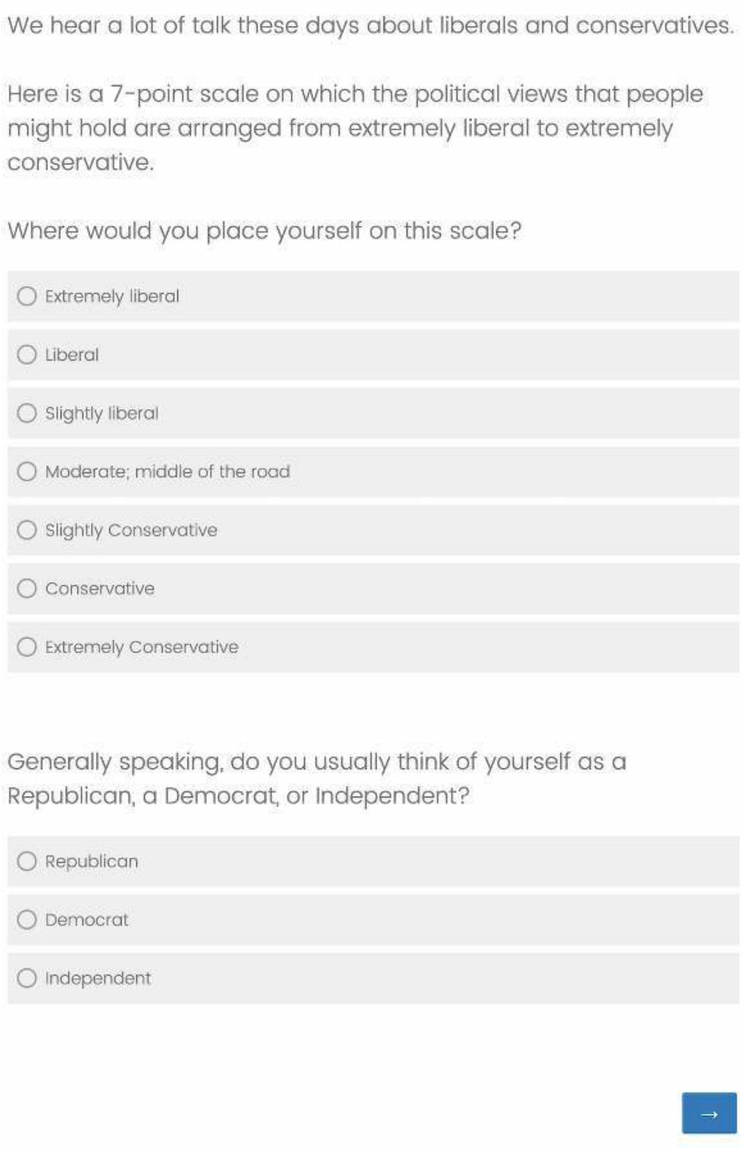
□ Native American or Alaska Native

□ Native Hawaiian or Pacific Islander

□ Prefer to self-describe:

Continued on next page

Appendix Table D1 – continued from previous page



We hear a lot of talk these days about liberals and conservatives.

Here is a 7-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative.

Where would you place yourself on this scale?

○ Extremely liberal

○ Liberal

○ Slightly liberal

○ Moderate; middle of the road

○ Slightly Conservative

○ Conservative

○ Extremely Conservative

Generally speaking, do you usually think of yourself as a Republican, a Democrat, or Independent?

○ Republican

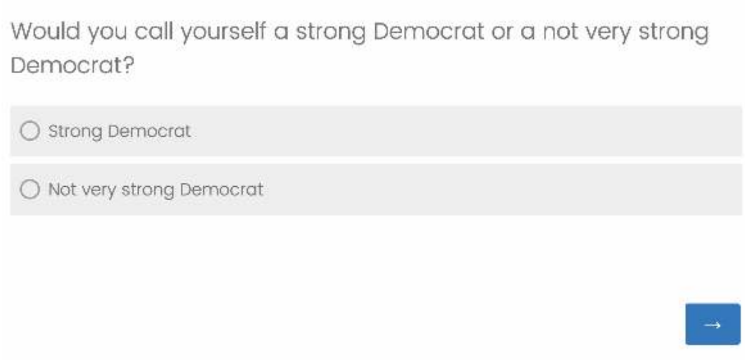
○ Democrat

○ Independent

Political identity (Democrat branch):

Continued on next page

Appendix Table D1 – continued from previous page



Would you call yourself a strong Democrat or a not very strong Democrat?

○ Strong Democrat

○ Not very strong Democrat

D.2 Instructional Videos

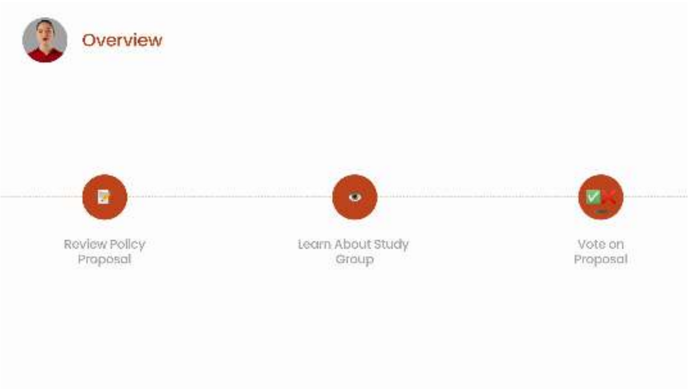
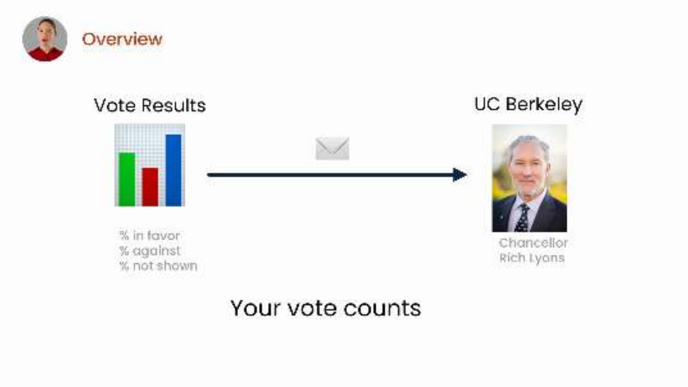
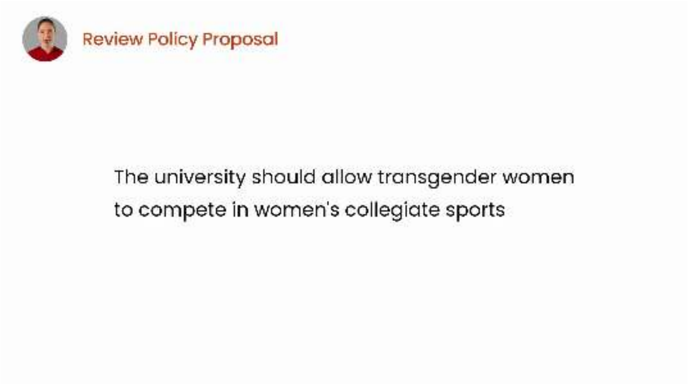
Tables D2 through D6 present the final frames of each slide and the corresponding voice-over transcripts from the instructional videos.

The tables show in order the shared instructions about the voting task, and then, the instructions specific to Private, Public, and Threshold Treatment. Table D6 shows the second stage instructional video specific to the Threshold treatment.

Appendix Table D2: Screenshots and transcripts from the instructional video: Introduction

Screen	**Voice-over Transcript**
	Screen 1: Welcome to the study *"Welcome to the study. You are invited to participate in a vote on an important campus policy proposal."*

Continued on next page

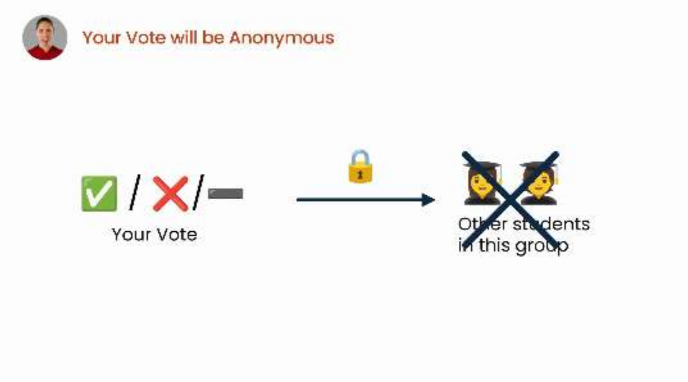
Screen	**Voice-over Transcript**
 Overview Review Policy Proposal Learn About Study Group Vote on Proposal	Screen 2: Overview *"Here's what will happen: First, you will review a specific policy proposal. Second, you receive information about the visibility of your vote to other students. Third, you vote in favor or against the proposal, or choose to abstain."*
 Overview Vote Results % in favor % against % not shown UC Berkeley Chancellor Rich Lyons Your vote counts	Screen 3: Vote counts *"The results of the vote — that is, the percentage of students choosing each option — will be shared with Chancellor Rich Lyons. This is a real opportunity to voice your opinion on a campus policy matter. Your vote counts."*
 Review Policy Proposal The university should allow transgender women to compete in women's collegiate sports	Screen 4: Your proposal *"Your policy proposal is the following. Please read it carefully."*

Continued on next page

Appendix Table D2 – continued from previous page

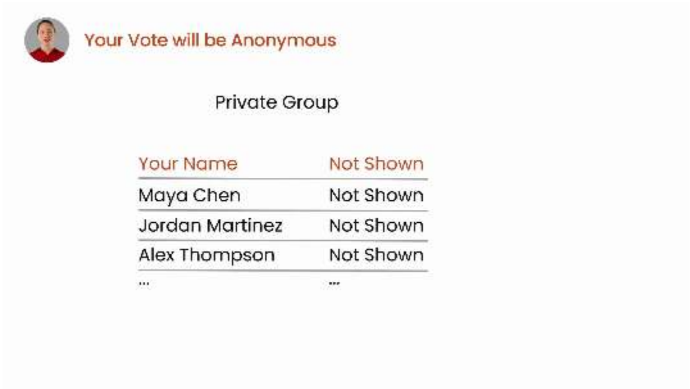
Screen	**Voice-over Transcript**
	Screen 5 [after treatment-specific slides]: End. *"Now it's time for you to vote!"*

Appendix Table D3: Screenshots and transcripts from the instructional video: Private Treatment.

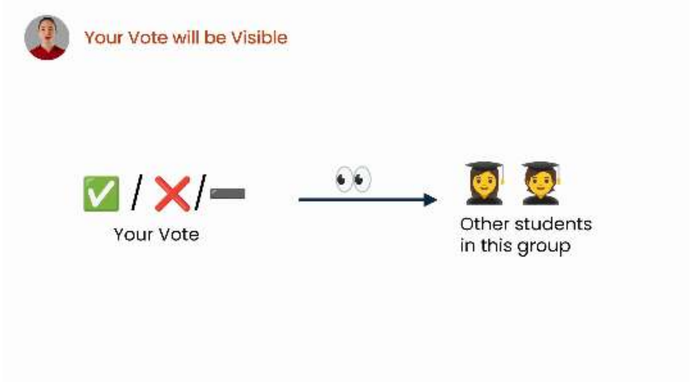
Screen	**Voice-over Transcript**
 Your Vote will be Anonymous ✅ / ❌/➖ Your Vote Other students in this group	Screen 1: Visibility *"In this group, your vote will never be linked to your name or identity. Your vote remains completely private."*

Continued on next page

Appendix Table D3 – continued from previous page

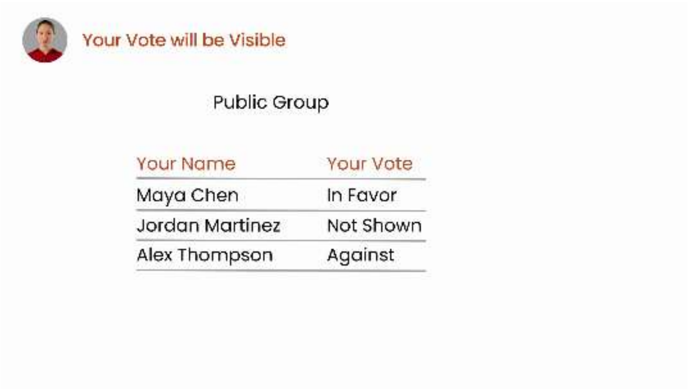
Screen	**Voice-over Transcript**
 Your Vote will be Anonymous Private Group Your Name — Not Shown Maya Chen — Not Shown Jordan Martinez — Not Shown Alex Thompson — Not Shown ... — ...	Screen 2: Sharing *"After the study ends, XLab staff will send out a spreadsheet showing the names of all participants in this group."*

Appendix Table D4: Screenshots and transcripts from the instructional video: Public Treatment.

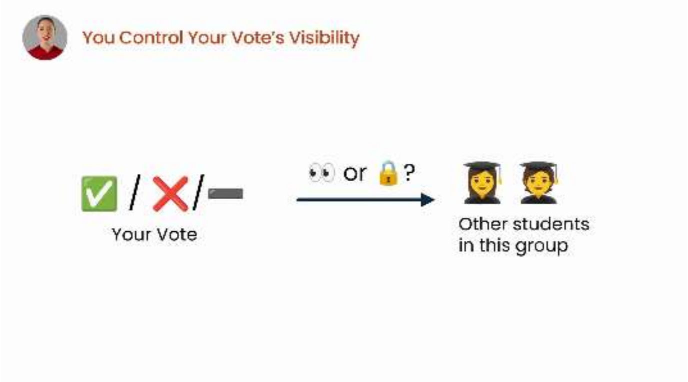
Screen	**Voice-over Transcript**
 Your Vote will be Visible ✅ / ❌/➖ Your Vote 👀 → Other students in this group	Screen 1: Visibility *"Your vote will be visible to other UC Berkeley students in this group."*

Continued on next page

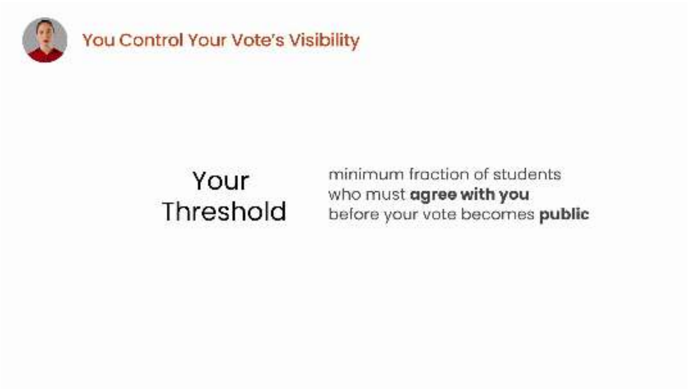
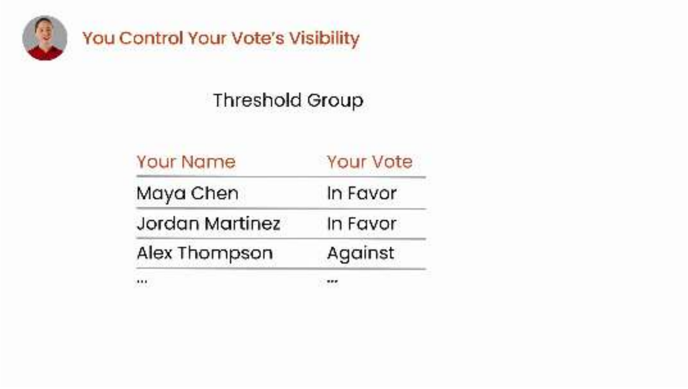
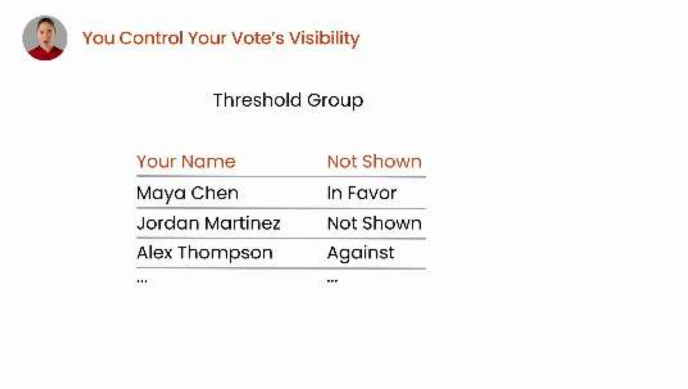
Appendix Table D4 – continued from previous page

Screen	Voice-over Transcript
 Your Vote will be Visible Public Group Your Name — Your Vote Maya Chen — In Favor Jordan Martinez — Not Shown Alex Thompson — Against	Screen 2: Sharing *"After the study ends, XLab staff will send out a spreadsheet showing how everyone in this group voted."*

Appendix Table D5: Screenshots and transcripts from the instructional video: Threshold Treatment.

Screen	Voice-over Transcript
 You Control Your Vote's Visibility ✅ / ❌/ ➖ Your Vote 👀 or 🔒? Other students in this group	Screen 1: Visibility *"In this group, you'll vote in two steps. First, you cast your vote anonymously. Then, you decide whether to share your vote with other students. You do this by setting a threshold."*

Continued on next page

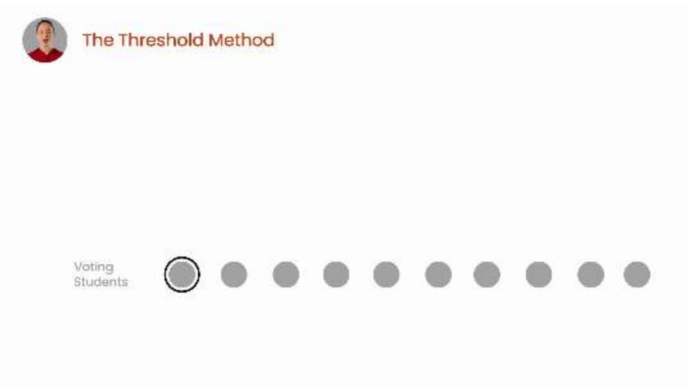
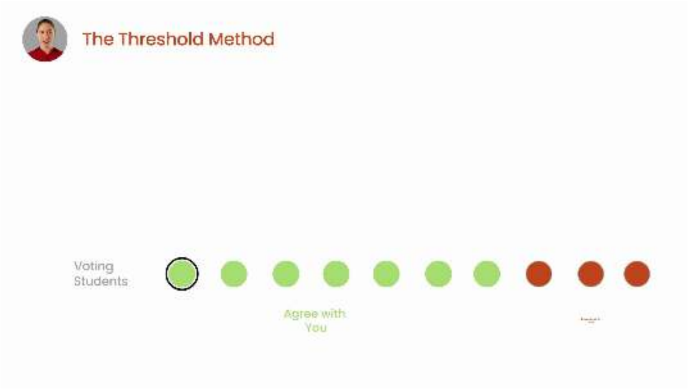
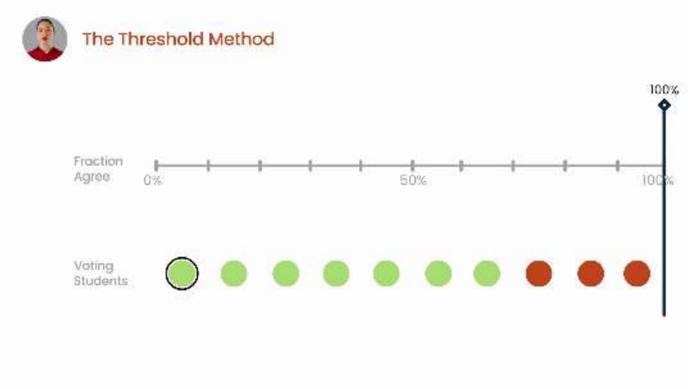
<table>
<tr><th>Screen</th><th>Voice-over Transcript</th></tr>
<tr><td>
You Control Your Vote's Visibility
Your Threshold
minimum fraction of students who must **agree with you** before your vote becomes **public**</td><td>Screen 2: Threshold
“The threshold is the minimum fraction of students who must agree with you before your vote becomes public. We’ll explain exactly how this works after you cast your vote.”</td></tr>
<tr><td>
You Control Your Vote's Visibility
Threshold Group<table><tr><th>Your Name</th><th>Your Vote</th></tr><tr><td>Maya Chen</td><td>In Favor</td></tr><tr><td>Jordan Martinez</td><td>In Favor</td></tr><tr><td>Alex Thompson</td><td>Against</td></tr><tr><td>...</td><td>...</td></tr></table></td><td>Screen 3: Sharing I
“After the study ends, XLab staff will send out a spreadsheet showing the names of all participants in this group. If you choose to share your vote, it will appear next to your name.”</td></tr>
<tr><td>
You Control Your Vote's Visibility
Threshold Group<table><tr><th>Your Name</th><th>Not Shown</th></tr><tr><td>Maya Chen</td><td>In Favor</td></tr><tr><td>Jordan Martinez</td><td>Not Shown</td></tr><tr><td>Alex Thompson</td><td>Against</td></tr><tr><td>...</td><td>...</td></tr></table></td><td>Screen 4: Sharing II
“Otherwise, your name will appear without a vote.”</td></tr>
</table>

Appendix Table D6: Screenshots and transcripts from the instructional video: Threshold Instructions.

Screen	Voice-over Transcript
	Screen 1: Introduction *"So far, we have asked you to vote anonymously. Now, you can choose whether to share your vote with other students. We plan for approximately 400 UC Berkeley students to participate."*
Take a public stand Connect with others Encourage open dialogue	Screen 2: Motivation *"You might choose to share your vote to take a public stand, connect with others, or encourage open dialogue."*
	Screen 3: Sensitivity *"However, we understand that the issue may be sensitive. That's why we're testing a new method to give you control over your vote's visibility: The Threshold Method. Here's how it works."*

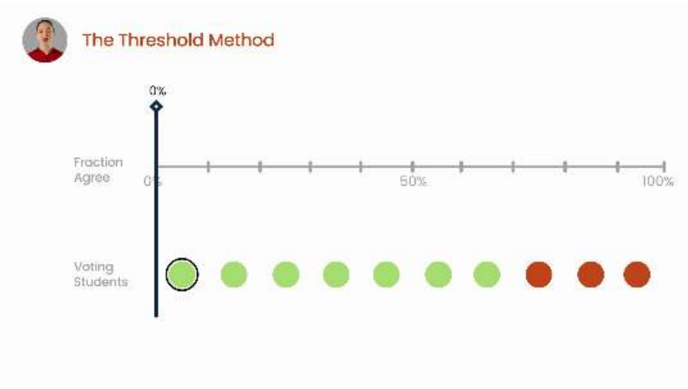
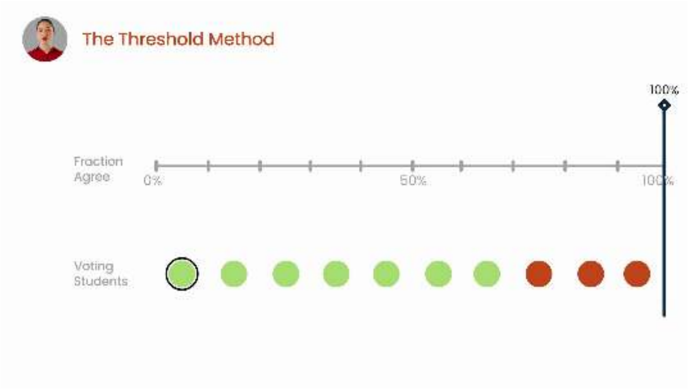
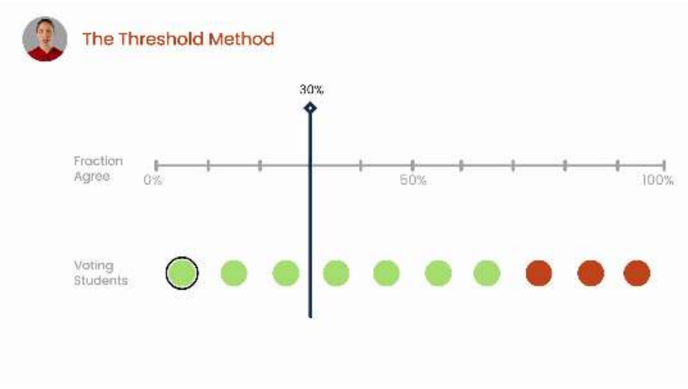
Continued on next page

Appendix Table D6 – continued from previous page

Screen	**Voice-over Transcript**
	Screen 4: Example Setup *"Let's consider a simple scenario. Here we have 10 students participating in a vote about a policy proposal. One of these students is you."*
	Screen 5: Voting Example *"Each student has voted either 'For' or 'Against' the proposal. In this example, 7 students voted the same way as you (shown in green) and 3 students voted differently from you (shown in red). You now need to decide whether to share your vote publicly, or keep it private."*
 The Threshold Method 100% Fraction Agree 0% 50% 100% Voting Students	Screen 6: Threshold Concept *"This is where the threshold method comes in. Your threshold is a percentage you choose. It's the minimum fraction of students who must agree with you before your name is connected to your vote."*

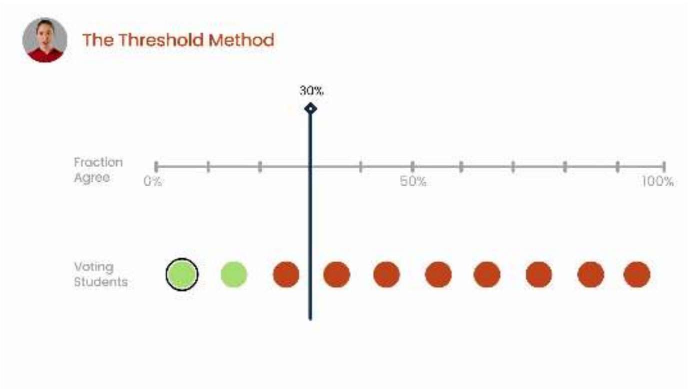
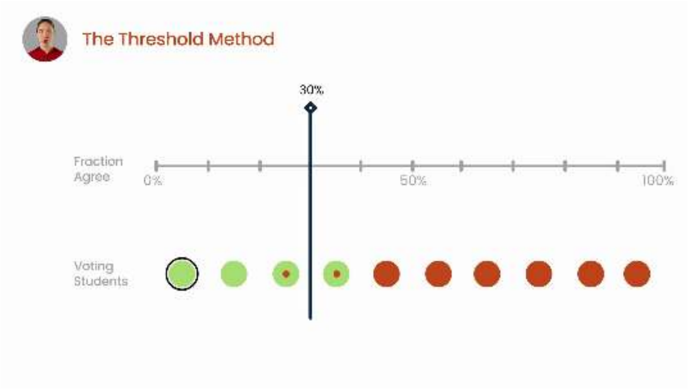
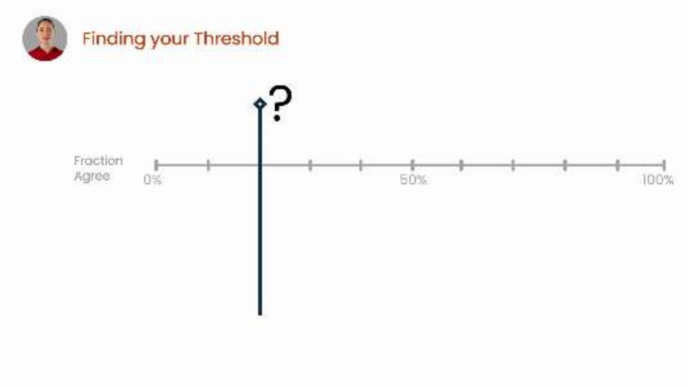
Continued on next page

Appendix Table D6 – continued from previous page

Screen	Voice-over Transcript
 The Threshold Method 0% Fraction Agree 0% 50% 100% Voting Students	Screen 7: 0% Threshold *"At 0%: Your vote is ALWAYS public. Your name appears with your vote no matter what."*
 The Threshold Method 100% Fraction Agree 0% 50% 100% Voting Students	Screen 8: 100% Threshold *"At 100%: Your vote is NEVER public. Your vote always stays hidden."*
 The Threshold Method 30% Fraction Agree 0% 50% 100% Voting Students	Screen 9: 30% Threshold *"Now let's try 30%: It now depends how others voted."*

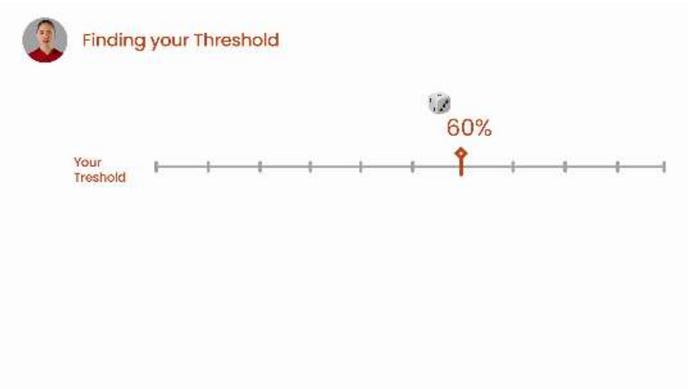
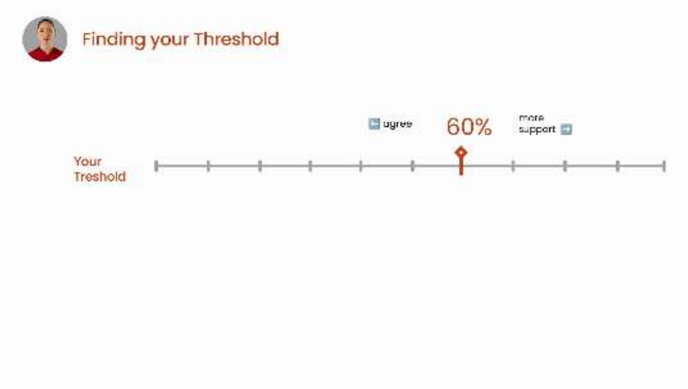
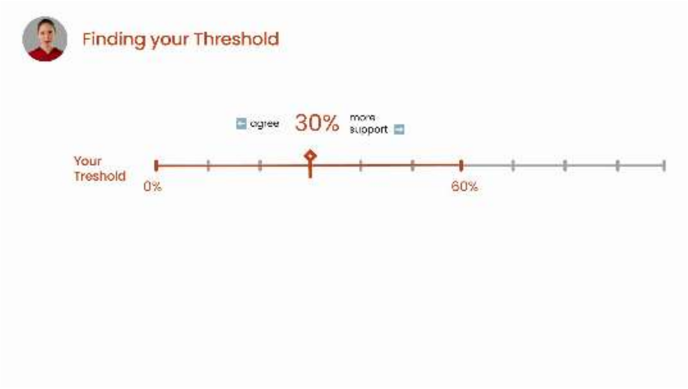
Continued on next page

Appendix Table D6 – continued from previous page

Screen	**Voice-over Transcript**
 The Threshold Method 30% Fraction Agree 0% 50% 100% Voting Students	Screen 10: Below Threshold *"Here, only 20% voted like you. That's below your 30% threshold. Your vote stays private."*
 The Threshold Method 30% Fraction Agree 0% 50% 100% Voting Students	Screen 11: Above Threshold *"Here, 40% voted like you. That's above your 30% threshold. Your vote becomes public. This is 'safety in numbers.' You decide exactly how much agreement you need to feel comfortable to share your vote."*
 Finding your Threshold ? Fraction Agree 0% 50% 100%	Screen 12: Step-by-Step *"To find your ideal threshold, we use a simple step-by-step process."*

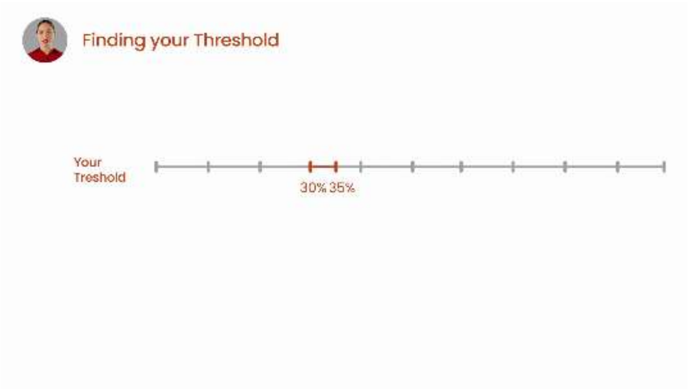
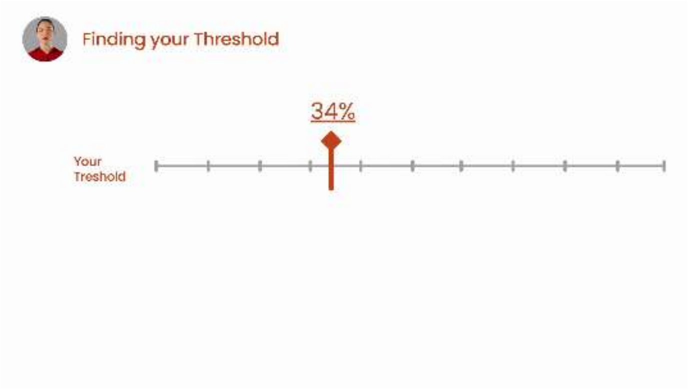
Continued on next page

Appendix Table D6 – continued from previous page

Screen	**Voice-over Transcript**
	Screen 13: Random Start *“We’ll start with a random threshold percentage.”*
 Finding your Threshold agree 60% more support Your Treshold	Screen 14: Comfort Check *“You’ll tell us if you are comfortable with that level of agreement, or if you need more students to agree with you.”*
 Finding your Threshold agree 30% more support Your Treshold 0% 60%	Screen 15: Adjustment *“Based on your answer, we’ll adjust and ask again.”*

Continued on next page

Appendix Table D6 – continued from previous page

Screen	**Voice-over Transcript**
	Screen 16: Narrowing Down *"After a few rounds, we'll narrow it down to a small range."*
	Screen 17: Final Choice *"Then you'll select your exact preferred threshold."*
	Screen 18: Practice Round *"Let's start with a practice round, then we'll find your threshold."*