Threshold Mechanism: Pre-Analysis Plan

October 2025

1 Introduction

This pre-analysis plan outlines a randomized online experiment testing a threshold voting mechanism that aims to balance privacy and information revelation in collective decision-making.

In many collective decisions, especially those concerning sensitive topics, individuals face competing pressures. Public voting fosters coordination and accountability but risks conformity through social image concerns. Private voting protects truthful expression but obscures the distribution of preferences. Standard voting systems thus force a stark trade-off: truthful voting at the expense of accountability, or full accountability at the expense of truthful expression.

We study an intermediate institution, the threshold voting mechanism. Individuals first cast their vote anonymously, then choose a personal threshold: their vote is revealed publicly only if the number of others voting the same way reaches or exceeds that threshold. This mechanism operationalizes the idea of 'safety in numbers.' From a theoretical standpoint, it preserves truthful voting as under anonymity while still disclosing information about the distribution of preferences.

In our experiment, UC Berkeley students enrolled in Fall 2025 are randomly assigned to private, public, or threshold voting conditions. Participants vote on a contentious policy proposal, with thresholds elicited only in the threshold treatment. This design allows us to measure (i) the distortions created by public voting and (ii) whether the threshold mechanism reduces those distortions while preserving information about underlying preferences.

2 Experimental Design

After providing informed consent, participants learn they will be asked to vote on whether the university should allow transgender women to compete in women's collegiate sports. They are informed that (1) aggregate voting results will be shared with the university chancellor, and (2) individual votes may be shared with other study participants, depending on their assigned treatment condition.

Participants will be randomly assigned to one of three treatments:

1. **Public Treatment:** Participants are informed that their vote and name will be shared with other participants. They then cast their vote (in favor, against, or abstain). If a subject chooses to abstain, we will still share her name with other participants and we will report her vote as "abstained."

2. **Private Treatment:** Participants are informed that their vote is anonymous and will not be revealed to other participants. They then cast their vote (in favor, against, or abstain). We will share the names of subjects assigned to the private treatment with other participants, but we will reveal neither whether they abstained, nor what they voted for in case they did not abstain.

3. **Threshold Treatment:** Participants are informed we will use a threshold mechanism for conditional vote revelation. They first cast their vote (in favor, against, or abstain). If they do not abstain, they are asked set a threshold (0-100%) specifying the minimum fraction of others who must vote the way they did in order for their vote to be revealed. For example, a 60% threshold means the vote becomes public only if at least 60% of participants not abstaining voted the same way. Participants complete a practice round and comprehension checks before setting their threshold.

Following the voting decision, all participants privately report their perceptions of the social acceptability of voting in favor or against the proposal, their own degree of engagement with the issue, and demographics including political ideology.

After data collection, results are distributed via email to participants. The email lists all participant names within their experimental group. In the Public Treatment, names appear with votes (except abstentions, where only the name appears). In the Private Treatment, only names appear. In the Threshold Treatment, names appear for everyone and votes appear only for participants whose thresholds were met.

3 Outcome Variables, Hypotheses, and Statistical Tests

For each participant, we collect vote choice (favor/against/abstain) and, in the threshold treatment only, their chosen threshold (0-100%). Vote revelation follows from treatment design: no votes revealed in private, all votes that are not abstentions revealed in public, and threshold-contingent revelation in the threshold treatment.

3.1 Preconditions for Threshold Mechanism to be Relevant

For our threshold mechanism to address the shortcomings of both private and public voting, public voting must be a distorted version of private voting. There-

fore, our aim is to recruit participants for the private and public treatments until the two margins below are significant, as determined by one-sided t-tests.

1. Abstention: We expect abstention rates to be higher under public than private voting and will test accordingly.

2. Expression of non-socially desirable views: Among all participants, we expect the share voting against the socially desirable position (i.e., the share of participants that do not abstain and vote for the non-socially desirable position) to be lower under public than under private voting and will test accordingly.

3.2 Main Outcomes & Hypotheses

We test two main hypotheses about the threshold mechanism:

1. Abstentions: We expect abstention rates to be higher under public than threshold voting. We will test accordingly using a one-sided t-test.

2. Expression of non-socially desirable views: We expect the share expressing non-socially desirable views (i.e., the share of participants that do not abstain and vote for the non-socially desirable position) to be higher under threshold than under public voting. We will test accordingly using a one-sided t-test.

3.3 Secondary Outcomes & Hypotheses

We expect the fraction of participants whose individual votes other than abstention are publicly revealed to be higher in the threshold mechanism than in the private mechanism (which is mechanically zero).

4 Sampling Strategy

4.1 Sample Size and Sequential Design

We use a sequential design with up to two stages.

Stage 1: We plan to exhaust the pool of participants at the UC Berkeley economics laboratory. The UC Berkeley lab manager estimates that the pool comprise around 1200 individuals, which would give us 400 individuals per treatment arm. Upon exhausting the pool of participants at the UC Berkeley economics laboratory, we will test whether the two preconditions described in Section 3.1 are met using t-tests with an α level of 0.05. If the two preconditions show significant differences, we will stop the data collection and analyze our main and secondary outcomes.

Stage 2: If the Stage 1 criteria are not met, we will calculate the sample size required for those preconditions to be met based on the observed effects

(targeting 80% power). We will then expand recruitment to a comparable university such as UC Santa Barbara. If the Stage 1 criteria are not met, we will upload an addendum this pre-analysis plan before collecting Stage 2 data.

The main treatment effects will be analyzed only after preconditions are satisfied. The stopping rule ensures the relevance of our proposed mechanism.

4.2 Robustness and Quality Flags

We exclude participants who fail a basic instruction manipulation attention check (e.g., "select agree").

Following instructional videos, all participants complete comprehension checks about the voting task. Participants who fail receive a summary of key instructions and are asked to retry. Performance may inform exploratory heterogeneity analyses (e.g., comparing behavior among those who understood immediately vs. after review) and we might include comprehension fixed effects in our analysis.

Participants in the threshold treatment complete additional comprehension questions about the mechanism. Performance may inform exploratory heterogeneity analyses (e.g., comparing behavior among those who understood immediately vs. after review) and we might include comprehension fixed effects in our analysis.

4.3 Heterogeneous Effects

We explore heterogeneity in treatment effects along i) self-reported engagement with the issue, ii) political ideology, iii) gender and sexual orientation.