Toward an Understanding of the Political Economy of Using Field Experiments in Policymaking

Pre-Analysis Plan

Guglielmo Briscese* John List†

May 2026

Abstract

This pre-analysis plan (PAP) specifies the data collection and empirical strategy for an additional survey experiment that complements findings from Briscese and List (2024). The new experiment replicates the core policymaker survey design using a different study as stimulus—a published field experiment testing the efficacy of social norms messaging on retirement savings contributions (Beshears et al., 2015)—and adds three novel elements: (i) a treatment arm providing explicit information about the methodological reliability of the study, designed to disentangle genuine evaluation aversion from credibility discounting; (ii) a module eliciting policymakers' anticipated public backlash as a mechanism; and (iii) a high-stakes incentive-compatible revealed preference measure, in which respondents allocate lottery tickets between a policy evaluation research partnership and a charitable donation of equivalent value. Both stated preferences for funding experimental policy evaluations and revealed preferences via the lottery allocation are designated as primary outcomes. Reputation and accountability perceptions are pre-specified as exploratory mechanisms.

Keywords: Evidence-based policymaking; beliefs; experimentation aversion; accountability; replication

JEL Classification: C93, D72, D83

*Harris School of Public Policy, University of Chicago. Email: gubri@uchicago.edu
†Department of Economics, University of Chicago. Email: jlist@uchicago.edu

1 Introduction

This pre-analysis plan outlines the design of a survey experiment on U.S. state policymakers that complements and extends the findings of Briscese and List (2024). In the previously published study, we explored how policymakers and a representative sample of the U.S. population react after being shown results from a large-scale natural field experiment that yielded null effects for a widely implemented policy incentivizing enrollment in 529 college savings accounts. The paper finds that policymakers, initially overoptimistic about program effectiveness, adjust their views based on evidence but show reduced demand for experimentation—suggesting *experiment aversion* when results defy expectations.

The present study pursues two main goals. First, it **unpackstwo mechanisms** that may drive the aversion result: experimentation aversion and public backlash. A key identification challenge is that policymakers who see disappointing results may reduce their demand for further experimentation either because (a) they *update too broadly*, generalizing from one null result to skepticism about the scientific enterprise as a whole—what we call genuine *evaluation aversion*—or (b) they *discount the reliability* of the specific study, treating the null result as uninformative rather than conclusive. To separate these channels we implement a three-arm design: a control group, a treatment group that sees the study results, and a second treatment group that sees the results alongside explicit information about the methodological quality and reliability of the study. If aversion is driven by credibility discounting, the reliability information should attenuate it; if aversion persists even when credibility is established, the result supports genuine evaluation aversion.

The additional mechanism we investigate is whether policymakers anticipate **public backlash**. Policymakers may curtail their demand for experimentation not because they personally oppose it, but because they expect that publishing disappointing results would harm their organization's public reputation and, in turn, threaten their careers. We measure this channel by eliciting policymakers' perceptions of potential public reactions and reputational consequences for their organization if a study like the one they read about were implemented and made public.

The second goal of the study is to introduce a **high-stakes incentive-compatible revealed preference measure**. All respondents allocate lottery tickets between two options: collaborating with University of Chicago faculty on a rigorous policy evaluation of their choice, or a $1,000 charitable donation made on their behalf. Both options are of comparable monetary value and the allocation is incentive-compatible, as one ticket is drawn at random to determine the prize received by the winner.

2 Survey Design

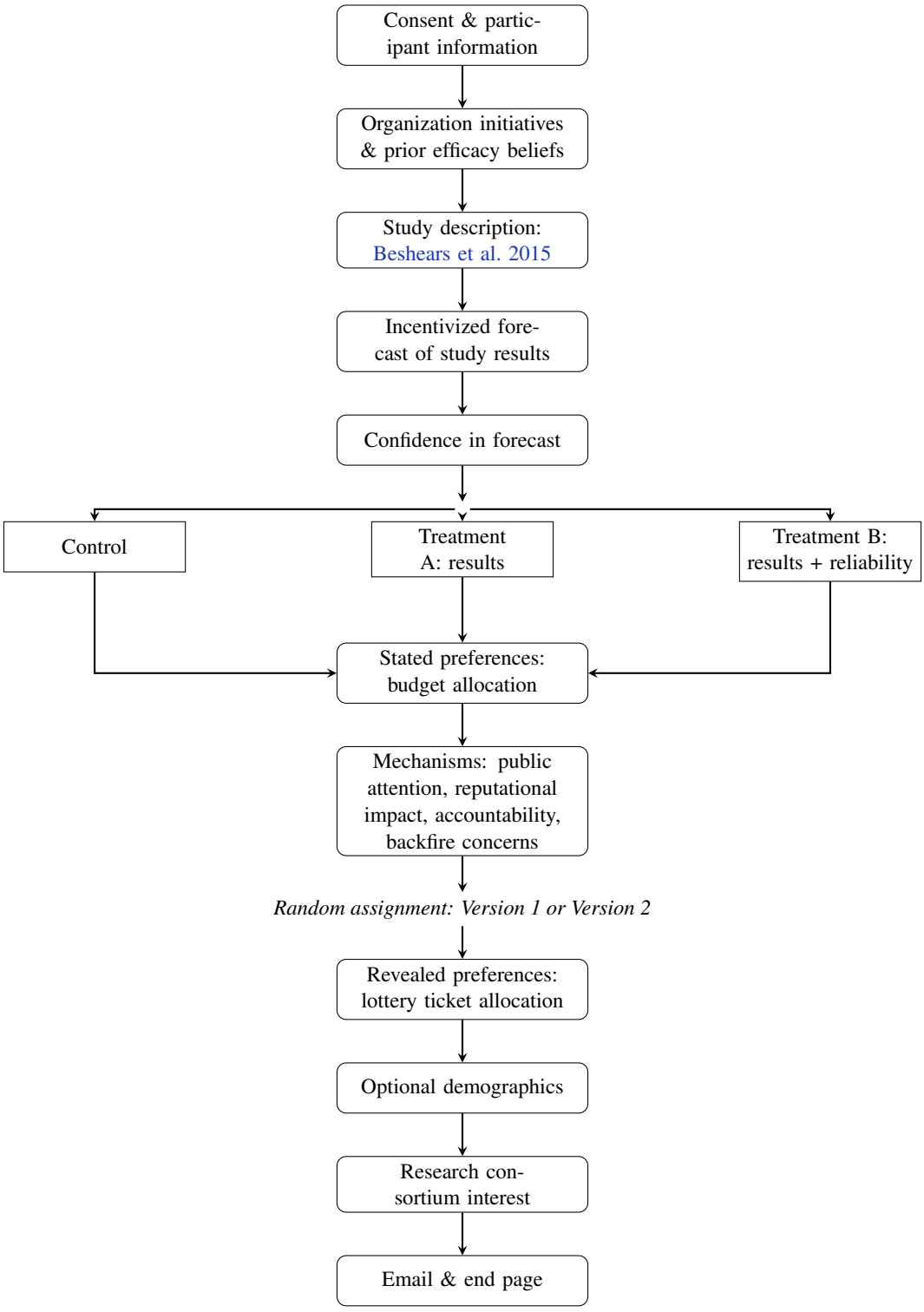
2.1 Sample and Recruitment

We recruit a sample of policymakers from U.S. state government agencies and contracted service providers responsible for the administration of savings programs, in line with the original study. The target population includes administrators of 529 college savings accounts, ABLE (Achieving a Better Life Experience) savings accounts, state-administered retirement savings programs, and related financial literacy programs, among others. We employ a convenience recruitment strategy covering both in-person and online channels over a period of a few months, targeting a sample of between 100 and 300 completed responses. As in the first survey, respondents are recruited via email invitations sent to agency staff through the National Association of State Treasurers in relation to a national conference scheduled in June 2026. In the previous study, we limited our recruitment strategy to an email invite and one follow-up, but for this additional survey we expect to send additional reminders as well as leverage the in-person conference in June, as needed.

2.2 Survey Flow

Figure 1 presents the survey flow. The survey is hosted on Qualtrics, with treatment randomization handled automatically by the platform's built-in block randomizer, which ensures even presentation across arms. Respondents are assigned with equal probability to one of three experimental conditions. Separately, the revealed preference question is randomized between two versions to account for ordering effects (see Section 2.3).

Figure 1: Survey flow



2.3 Survey Modules

Module 1: Consent. Respondents read a participant information statement and confirm their willingness to participate before proceeding.

Module 2: Organizational initiatives and prior efficacy beliefs. Respondents first indicate which types of outreach initiatives their organization typically implements to encourage savings—such as generic email campaigns, social norms messaging, and seeding incentives. They then rate the perceived effectiveness of each type of initiative on a 5-point scale. This module anchors respondents' prior beliefs and captures their professional context before exposure to the study.

Module 3: Study description. Respondents read a brief, plain-language description of Beshears et al. (2015), a field experiment that mailed letters to employees enrolled in a 401(k) plan but contributing below the level required to receive the full employer match. Recipients were randomly assigned to one of three conditions: no information about peer savings behavior; a sentence reporting the savings behavior of coworkers in the same five-year age bracket; or the same sentence for a ten-year age bracket. The study measured whether recipients subsequently increased their contribution rate. None of the three groups differed significantly from the others, demonstrating that social norms messaging did not change employees' savings behavior.

We selected this study for three reasons. First, social norms communications are widely used in savings promotion, so most respondents will have direct professional experience with similar outreach. Second, the study is a well-designed randomized experiment published in the *Journal of Finance*—the highest-ranked peer-reviewed journal in household finance—which supports our mechanism test regarding credibility discounting. Third, it is among a small number of null-result studies to appear in a top-tier economics journal, providing a credible and ecologically valid stimulus. Respondents are not shown the study results at this stage.

Module 4: Incentivized forecast. As in Briscese and List (2024), we elicit respondents' expectations about the study's results in an incentive-compatible manner. To minimize cognitive burden, respondents are shown the number of employees in each randomized group and asked to estimate how many in each group subsequently increased their contribution rate. Respondents whose forecasts fall within $\pm 30\%$ of the actual values are entered into a drawing for a \$50 Amazon gift card. Respondents then rate their confidence in their estimates on a 7-point scale. In the previous study, we used a \$25 voucher because

we elicited respondents' forecasts twice, thus giving every participant the opportunity to earn up to $50 in survey completion rewards. Since this survey is shorter, and we only employ one forecasts elicitation question, we increased the voucher amount to $50 for comparability and higher recruitment purposes.

Module 5: Treatment randomization. Following the forecast module, respondents are randomly assigned with equal probability to one of three conditions:

- **Control:** Respondents see only a thank-you message for providing their forecast and are prompted to continue.
- **Treatment A (results only):** Respondents see the actual results of the study, presented in the same format as the forecast task—the number of employees who increased their contribution rate in each group—with no additional commentary.
- **Treatment B (results + reliability information):** Respondents see the same results as Treatment A, together with a brief set of bullet points explaining the methodological quality and reliability of the study. This information covers the study's randomized design, its sample size and statistical power, its publication in the *Journal of Finance*, and why null results in well-powered studies provide informative evidence. A link to the full paper is also provided, and whether respondents click on it is recorded as a behavioral indicator of engagement with the evidence. The purpose of this arm is to hold the informational content about results constant while varying the degree to which respondents are encouraged to treat the evidence as credible.

The key theoretical contrast is between Treatment A and Treatment B: if evaluation aversion is driven by credibility discounting, Treatment B should attenuate the effect; if aversion persists at a similar magnitude in both treatment arms, the result is consistent with genuine aversion to experiments that produce unwelcome findings.

Module 6: Stated preferences—budget allocation. Consistent with Briscese and List (2024), all respondents complete a budget allocation task in which they distribute a hypothetical organizational budget of $100,000 across four options, presented in randomized order:

1. Implement a new, different trial testing a completely different intervention to increase savings;
2. Apply the same social norms messaging to other communication or marketing campaigns (i.e., scale up the intervention);
3. Increase funding for business-as-usual programs and initiatives;

4. Replicate the same trial with methodological changes (e.g., different recipient cohorts, larger sample size).

Respondents enter a percentage share for each option, with the constraint that all four shares sum to 100. This question is a **primary outcome** of the study.

Module 7: Mechanism module. This module contains four question batteries eliciting potential mechanisms underlying treatment effects on experimentation demand.

- **Public attention:** Respondents indicate, on a 7-point scale, the degree to which they expect the general public would be interested in learning about a study like the one they read about, if it were conducted and made public by their organization.
- **Reputational impact:** Respondents rate, on a 7-point scale ranging from very negative to very positive impact, how conducting and publishing such a study would affect their organization's reputation with the public along three dimensions: ability to do its job well; care for the people it serves; and honesty and trustworthiness.
- **Accountability:** Respondents rate, on a 7-point scale, the extent to which they personally feel accountable to each of five stakeholder groups—state legislators and oversight committees; the general public; program beneficiaries; their manager or organizational leadership; and professional peers in other states or organizations—for the outcomes of programs their organization implements.
- **Backfire concerns:** Respondents rate their level of concern, on a 7-point scale, about four potential negative consequences if a program their organization administers were evaluated and found to be ineffective: negative media coverage; public criticism; budget cuts to their programs; and damage to their professional reputation. These four items are combined into a **composite backfire concern index** (simple mean), which is the pre-specified outcome for this mechanism. Internal reliability (Cronbach's α) will be reported. If $\alpha < 0.60$, effects on individual items will be reported in place of the composite.

Module 8: Revealed preferences—lottery ticket allocation. All respondents are informed that they will be entered into a random draw as a token of appreciation. One randomly selected respondent will win one of two prizes, determined by their ticket allocation. The two prize options are:

- **Policy Evaluation Research Partnership:** University of Chicago faculty will work with the winner at no cost to co-design and implement a rigorous experimental evaluation of a program or

policy of their choice, at a time and pace that works for them, with results published in academic and policy outlets.

- **Charitable Donation:** A $1,000 donation will be made on the winner's behalf to United Way, a nonprofit that funds local organizations working to reduce poverty across the United States.

Respondents allocate 100 lottery tickets between the two options. Rather than asking for a binary choice, the ticket allocation allows respondents to express how strongly they prefer one option over the other: if selected as the winner, one ticket is drawn at random to determine which prize they receive. This design is incentive-compatible with real opportunity cost, as the fractional allocation directly determines the probability of each outcome.

To control for potential order effects, respondents are randomly assigned to one of two versions of this question. In **Version 1**, the charitable donation option is listed first; in **Version 2**, the research partnership option is listed first. In addition, within each version, the order of the two numeric input boxes is independently randomized. Version assignment and input box order are recorded as embedded data variables and included as controls in the primary analysis (see Section 4). The primary outcome variable is the number of lottery tickets allocated to the research partnership, regardless of version or input box order.

Module 9: Demographics (optional). Respondents complete a brief optional demographic questionnaire covering: number of years with their current employer; expected years remaining in current role; highest level of education; age; and gender.

Module 10: Research consortium interest and end page. Respondents are asked whether they would like to receive more information about a national research consortium—a collaborative initiative between the University of Chicago and NAST to conduct rigorous evaluations of savings programs across states. Expressed interest in this consortium serves as an additional behavioral indicator of demand for policy experimentation. The survey concludes with a request for respondents' email addresses to (a) send prize instructions if selected as the lottery winner; and (b) prevent duplicate responses.

3 Hypotheses and Outcomes

This pre-registered experiment investigates four research questions. We state each as an open empirical question rather than a directional prediction, as the prior evidence does not unambiguously determine

the sign of the expected effects. We discuss the plausible mechanisms behind each possible direction.

RQ1 (Stated preferences): Does exposure to a published null-result study affect policymakers' stated preferences for how to allocate a budget across experimental evaluations, scale-up, and business-as-usual activities?

The direction of this effect is a priori unclear. On the one hand, disappointing results may trigger *evaluation aversion*, reducing policymakers' willingness to fund further experimental evaluations and shifting resources toward scaling existing programs or business-as-usual activities—consistent with the main finding in Briscese and List (2024). On the other hand, exposure to a rigorous evaluation study may trigger *experimenter demand effects*: policymakers, now primed to think about scientific evaluation, may report *higher* stated preferences for new experiments, regardless of whether the results were positive or null. A third possibility is that the effect is heterogeneous across the budget categories in ways that differ from the original study, given the different stimulus and respondent composition. We therefore treat this as an open empirical question and examine treatment effects on each budget allocation category separately.

RQ2 (Revealed preferences): Does exposure to a published null-result study affect policymakers' revealed preferences for engaging in policy evaluation, as measured by the incentive-compatible lottery ticket allocation between a research partnership and a charitable donation?

The revealed preference measure is designed to complement the stated preference outcome and provide a test that is possibly less susceptible to experimenter demand effects. If experimenter demand drives stated preferences upward, revealed preferences—which carry real opportunity cost and are less socially observable—may move in the opposite direction, or not move at all. Conversely, if genuine evaluation aversion is the dominant mechanism, we would expect both stated and revealed preferences to shift in the same direction. A divergence between the two primary outcomes would itself be informative about the nature of the response.

RQ3 (Mechanism — credibility discounting vs. genuine aversion): Do the effects observed in RQ1 and RQ2, if any, differ between Treatment A (results only) and Treatment B (results plus reliability information)?

This question is designed to distinguish between two mechanisms. If any treatment effect on stated or revealed preferences is driven by respondents *discounting the credibility* of the null result—treating it as a product of poor methodology rather than a genuine finding—then providing explicit information

about the study's methodological quality (Treatment B) should attenuate the effect relative to Treatment A. If, on the other hand, the effect persists at similar magnitudes across both treatment arms, the result is more consistent with a *genuine aversion* to experiments that produce unwelcome findings, one that is not resolved by reassurances about study quality. An absence of a significant difference between Treatment A and Treatment B on a given primary outcome activates the pre-registered pooling rule for that outcome (see Section 4.2).

RQ4 (Mechanism — anticipated public backlash): Does exposure to the study results affect policymakers' concerns about potential negative public reactions to a similar evaluation being conducted and published by their organization?

The direction of this effect is also not predetermined. Seeing a null result may heighten concern about public backlash—media criticism, budget cuts, or reputational damage—if policymakers believe the public will interpret a failed evaluation as evidence of organizational incompetence. Alternatively, policymakers may find the study results unremarkable or may believe that the public is unlikely to follow academic research closely, leaving concern levels unchanged. We test this question by undertaking descriptive exploratory analysis using data collected from the mechanisms questions.

4 Empirical Strategy

4.1 Primary Outcomes

The study has two primary outcomes: (i) **stated preferences**, measured by the share of the hypothetical $100,000 budget allocated to new experimental evaluations; and (ii) **revealed preferences**, measured by the number of lottery tickets allocated to the research partnership. Both outcomes are analyzed using the same empirical framework, with a pre-registered pooling rule applied independently to each.

We estimate treatment effects using the following baseline OLS specification, consistent with Briscese and List (2024):

$$Y_i = \beta_0 + \beta_1 \,\mathrm{TreatA}_i + \beta_2 \,\mathrm{TreatB}_i + \varepsilon_i \tag{1}$$

where Y_i is the outcome of interest for respondent i, TreatA_i and TreatB_i are indicators for assignment to Treatment A and Treatment B respectively, and the control group is the omitted category.

We additionally report the extended specification:

$$Y_i = \beta_0 + \beta_1 \,\mathrm{Treat\!A}_i + \beta_2 \,\mathrm{Treat\!B}_i + \mathbf{X}_i'\boldsymbol{\gamma} + \varepsilon_i \tag{2}$$

where $\mathbf{X}_i$ is a vector of pre-specified covariates: a dummy for whether the respondent's organization implements similar campaigns, a postgraduate education dummy, a female dummy, age group indicators, and job tenure indicators. For the revealed preference outcome, $\mathbf{X}_i$ additionally includes a dummy for question version (Version 1 vs. Version 2) and a dummy for input box order, to absorb any residual order effects.

4.2 Pre-registered Pooling Rule

The three-arm design serves two purposes: (i) the primary test of evaluation aversion, contrasting each treatment arm against control; and (ii) the mechanism test separating credibility discounting from genuine aversion, contrasting Treatment A against Treatment B. Pooling the two treatment arms when they do not differ increases statistical power on both primary outcomes.

We pre-register the following conditional pooling rule, applied **independently** to each primary outcome:

> *For each primary outcome, we first conduct a simple two-sample t-test comparing Treatment A against Treatment B on that outcome. If this comparison yields no statistically significant difference, the two treatment arms are pooled into a single combined treatment group for all analyses of that outcome involving the treatment vs. control contrast. The pooled specification is then reported as the main result for that outcome, with the three-arm specification reported alongside it for transparency. If a statistically significant difference between Treatment A and Treatment B is detected, the pooling rule is not applied for that outcome: the three-arm specification in Equation 1 remains primary, exactly as it would have been without this pooling amendment.*

The pooling decisions for stated and revealed preferences are made independently. It is therefore possible that the two treatment arms are pooled for one primary outcome but not the other, depending on the results of the respective t-tests. When pooling is applied to a given outcome, the primary estimating equation for that outcome becomes:

$$Y_i = \beta_0 + \beta_1 \,\mathrm{Treat}_i + \varepsilon_i \tag{3}$$

where $\text{Treat}_i = 1$ for any respondent assigned to Treatment A or Treatment B, and 0 for the control group. The extended specification with covariates follows analogously.

4.3 Revealed Preference Randomization

As described in Section 2.3, the lottery ticket question is randomly assigned to one of two versions differing in the order in which the prize options are presented. Within each version, the order of the numeric input boxes is independently randomized. These two randomizations are implemented via Qualtrics block randomization with even presentation, and the resulting assignment indicators are recorded as embedded data variables.

The primary analysis uses the number of tickets allocated to the research partnership as the dependent variable, regardless of version. As a pre-specified robustness check, we report results separately by version to confirm that the treatment effects are not driven by order effects. We also report results with and without version and input box order dummies included as covariates, as specified in Equation 2.

4.4 Mechanism Outcomes

The following outcomes are pre-specified as mechanisms and are analyzed using Equation 1. They are exploratory and not subject to multiplicity corrections.

- **Backfire concern questions:** Distributions of each of the four backfire concern items (negative media coverage, public criticism, budget cuts, professional reputation damage).
- **Reputational impact:** Three items capturing how publishing the study would affect the organization's perceived ability, care for beneficiaries, and honesty, analyzed separately.
- **Public attention:** A single item capturing expected public interest in the study if made public.
- **Accountability:** Five items capturing felt accountability to different stakeholder groups.
- **Research consortium interest:** A binary indicator for whether the respondent expressed interest in receiving more information about the national research consortium. Analyzed as an additional behavioral indicator of demand for policy experimentation, consistent with the approach in Briscese and List (2024).
- **Study link clicked:** A binary embedded data indicator for whether respondents in Treatment B clicked the link to the full paper. Analyzed as an indicator of engagement with the evidence.

4.5 Heterogeneous Effects

We will explore heterogeneous treatment effects by interacting treatment indicators with the following pre-specified moderators: (a) respondent's prior belief about the efficacy of outreach campaigns to measure possible disappointment effects; and (b) job tenure, as a proxy for ability to navigate organizational politics to implement a policy experiment. These analyses are exploratory and descriptive.

5 Power Calculations

Power calculations are benchmarked on the policymaker sample from Briscese and List (2024), using observed group means and standard deviations directly. For the stated preference outcome, the original study reports a control mean of 0.205 (SD 0.143) and a treatment mean of 0.309 (SD 0.255), with approximately 65 respondents per arm. For the revealed preference outcome, the original study reports a control mean of 0.675 (SD 0.290) and a treatment mean of 0.603 (SD 0.280). All power estimates are computed using Satterthwaite's t-test assuming unequal variances, and use a significance level of $\alpha = 0.10$ across all outcomes for comparability. For stated preferences we apply a two-sided test; for revealed preferences a one-sided test, consistent with their pre-specified directional role (Section 4).

Table 1 reports power estimates under two scenarios for each outcome: the "split" scenario, in which the control arm is compared to a single treatment arm (each of size $N/3$ under equal three-arm allocation), and the "pooled" scenario, in which the control arm ($N/3$) is compared to the combined treatment group ($2N/3$), activated by the pre-registered pooling rule in Section 4.2. For stated preferences, the split scenario at $N = 150$ already reaches 80% power (0.804), and pooling raises this substantially to 0.939. For revealed preferences, the two scenarios yield nearly identical power (0.805 in both cases) because the outcome has near-equal variance between control and treatment groups, meaning the efficiency gain from the unequal pooled allocation is negligible. Reaching 80% power for revealed preferences requires approximately $N = 290$ respondents under the split specification (145 per arm) or $N = 330$ under the pooled specification (110 control, 220 pooled treatment). At smaller sample sizes, effects on the revealed preference outcome should be interpreted as directional evidence consistent with the stated preference findings, in line with its pre-specified role as a validation measure.

Table 1: Power calculations benchmarked on Briscese and List (2024)

	Stated preferences		Revealed preferences	
	Split (N_1=50, N_2=50)	Pooled (N_1=50, N_2=100)	Split (N_1=145, N_2=145)	Pooled (N_1=110, N_2=220)
m_1 (control mean)	0.205	0.205	0.675	0.675
m_2 (treatment mean)	0.309	0.309	0.603	0.603
δ (effect)	+0.105	+0.105	−0.072	−0.072
σ_1 (control SD)	0.143	0.143	0.290	0.290
σ_2 (treatment SD)	0.255	0.255	0.280	0.280
Power	**0.804**	**0.939**	**0.805**	**0.805**

6 IRB Approval and Consent

The proposal has been approved by the University of Chicago Institutional Review Board: Approval No. IRB24-0796). All respondents provide informed consent prior to participation by reading and agreeing to the participant information statement at the start of the survey. Participation is voluntary and respondents may withdraw at any time without penalty. Data are collected anonymously; email addresses collected in the final module are stored separately from survey responses and used solely for lottery prize administration and deduplication.

References

Beshears, J., J. J. Choi, D. Laibson, B. C. Madrian, and K. L. Milkman (2015). The effect of providing peer information on retirement savings decisions. *The Journal of finance 70*(3), 1161–1201.

Briscese, G. and J. A. List (2024). Toward an understanding of the political economy of using field experiments in policymaking. Technical report, National Bureau of Economic Research.