The Political Economy of Evidence-based Policymaking

Guglielmo Briscese[*], John List[†]

October 14, 2024

Abstract

This study examines policymakers' and the general public's demand for evidence-based policies. Leveraging the results of a large-scale field experiment, we implement a set of survey experiments on a sample of U.S. state policymakers and a representative sample of Americans to investigate whether support for robust policy evaluation and scaling is influenced by prior beliefs on efficacy of the policy, and how respondents update their beliefs and preferences when presented with novel experimental evidence.

Keywords: Evidence-based policymaking; beliefs; experimentation aversion

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

The use of rigorous evaluation methods to assess the impact of public policies is gaining increased attention, as these approaches are vital for guiding the implementation and scaling of evidence-based interventions. However, despite their growing recognition, experimental evaluations and similar methods are still underutilized in public policy, both in the U.S. and globally. They are often limited to small-scale, newly introduced interventions, rather than being applied to evaluate large-scale programs that have been in place for several years. Furthermore, recent evidence indicates that even when interventions are robustly evaluated and proven effective, they often fail to be scaled up due to institutional barriers. This highlights the crucial role policymakers play in deciding which programs to adopt or expand.

Despite the important role policymakers play in adopting evidence-based policies, we know surprisingly little about their beliefs and preferences. A small body of research suggests that policymakers' beliefs about program effectiveness can be influenced, that there is demand for experimentation, and that they are willing to replicate successful interventions when organizational frictions are low and political alignment is present (Garcia-Hombrados et al., 2024; Toma and Bell, 2024; Vivalt and Coville, 2023; Hjort et al., 2021; DellaVigna et al., 2022). This study adds to this emerging literature by investigating policymakers' beliefs and preferences through a survey experiment. We also provide novel experimental evidence on citizens' demand for evidence-based policymaking and how this affects their trust in public institutions.

We draw on a pre-registered, two-phase randomized controlled trial (RCT) conducted in partnership with a government agency, which tested the efficacy of small financial incentives designed to boost enrollment in 529 College Savings Accounts (AEARCTR-0012055). While financial incentives are a common feature of many 529 plans across the U.S., there is limited evidence on their effectiveness. The trial revealed low uptake across all randomized groups, questioning the efficacy of such a long-standing program offered across several U.S. states.

To explore policymakers' reactions to this novel evidence, in May 2024 we conducted a survey experiment with attendees of a national conference of staff from state Treasurer's Offices responsible for administering 529 plans in each state to examine how they would respond to new evidence that may contradict their prior beliefs (unlike previous studies that only showed survey participants the trials re-

sults of successful interventions). The main findings from this survey experiment is that policymakers are overoptimistic about the efficacy of their policies and that exposure to experimental evidence that contradicts their prior beliefs leads to experimentation aversion.

In this pre-analysis plan, we list the hypothesis and empirical strategy of a complementary survey experiment on a representative sample of Americans. The goal of this pre-registered survey experiment is twofold: (i) assess whether, and how, the views of policymakers differ from those of the general public, and (ii) if policymakers' negative reaction to new evidence may be partly explained by the expectation that the general public would react negatively when learning about the evaluation results, potentially posing a risk to their careers. In line with the policymakers' survey experiment, we will randomly assign a subset of the general public respondents to view the results of the aforementioned trial before eliciting their preferences for public spending. In addition to the previous survey, we will also include a second randomized group that is shown the same results along with a brief text that aims to educate respondents on the importance of supporting robust evaluations in public policy, even when findings are unexpected. We will assess the effect of both treatment on support for trials to evaluate public policies and trust in public institutions.

2 Experiment and sample

In this section, we explain the survey experiment on the representative sample of the U.S. population. We replicate the main questions we asked in a previous survey on the sample of U.S. state policymakers, and add prior and posterior beliefs questions on support for policy experimentation and trust in institutions. The sample consists of N=1,200 respondents recruited via the Prolific survey platform. We will host the survey on Qualtrics, and randomization is introduced automatically using the survey software automatic randomizer. Figure 1 shows the survey flow.

The main difference between the general public and the previously implemented policymakers' survey is the introduction of a set of questions eliciting (a) their support for RCTs, and (b) their trust in public institutions, which we ask at the beginning and at the end of the survey, following the standard practice in survey experiments measuring changes in beliefs and policy preferences (Haaland et al., 2023).

The survey begins by measuring respondents' prior trust in public institutions - namely, their state treasurer's office, which is the agency responsible for administering the program they will be introduced to shortly after, and their state governor's office. The purpose of introducing both government agencies is twofold: we want to measure trust in different state institutions for which respondents may have different perceptions of politicization, with the treasurer's office likely being an institution respondents are less familiar with, and we also want to measure institutional trust spillover effects. We measure trust in institutions using a validated methodology that has been extensively used in management science and by public administration scholars, but comparatively less in the economics discipline. These studies use a method first proposed by Mayer (1995) which decomposes trust perceptions into three subdimensions: ability (i.e., organizational and personnel ability to deliver quality services to citizens), benevolence (i.e., motives and values), and integrity (i.e., a commitment to transparency and honesty).

Respondents are then asked to read a short paragraph explaining what experimental methodologies are and how they can be beneficial for improvements to public policies before being asked for their support for this evaluation approach. This module allows us to control and partly remove individual differences in prior knowledge on RCTs as well as partly control for experimenter demand effects: by showing this brief text to all respondents, we inform all survey participants of the technicalities and importance of using these methodologies in public policy.

The third module asks them their prior knowledge, experience, and beliefs about 529 plans. Regardless of their answers, all respondents see a brief explanation of what the plans are and how they may help families save for college so that the program goals are common knowledge. The goal here is again that of controlling for different prior knowledge of these plans that may influence subsequent responses.

At this stage of the survey, the modules are identical to the previously implemented policymakers' survey: respondents read a brief explanation of the pilot experiment testing the efficacy of small incentives to increase 529 plans take-up rates, and are asked in an incentive-compatible way to guess the pilot results. The incentive is such that respondents who forecast within -/+ 30% of the actual results will be entered into a drawing for a $25 Amazon gift card. We will also ask respondents how confident they are in their forecasts. To avoid biasing respondents' forecasts, we remove any information about the

trial partner name, just like we did for the policymakers' survey, and we ask participants to imagine the trial was implemented in their state. To control for whether respondents take this hypothetical framing seriously, we randomize respondents from Illinois into two subgroups: one group of Illinois respondents sees the same wording as all other respondents, while the other half of Illinois respondents sees the name of the trial partner agency.

Participants are then randomized into either one of three groups: a control group that only sees a thank you message for providing their best guess; a treatment group that see the results of the pilot experiment; and a third group (the only feature that differs from the policymakers' survey) where respondents see the pilot results complemented with a brief explanation on why such results are still important to inform evidence-based policymaking. The purpose of this additional trial arm is to see whether a light-touch scalable information provision intervention can partly counteract possible negative effects that may arise from showing respondents the results of the trial.

All respondents, just like in the policymakers' survey, complete a resource allocation task in which they will indicate how they would like their state government to allocate a hypothetical sum of $100,000 (a typical project budget in state governments) across multiple initiatives: replicating the seeding trial with a different audience sample, conducting the same trial with a larger sample, increasing funding for business-as-usual programs and initiatives, and conducting a new, different trial testing a different intervention. In a second task, participants must decide how to allocate a fixed amount of money between evaluating the effectiveness of a large-scale policy that has been in place for several years and evaluating the effectiveness of scaling up a pilot policy that has shown promising initial results. We then measure perceptions of the risk of spillovers and unintended consequences by asking participants how likely they think it is that different types of spillovers and unintended consequences will occur if the program being evaluated is scaled up. These questions aim to examine how respondents adjust their preferences for public spending upon learning the results of a trial, taking into account (unlike previous studies) possible trade-offs.

After measuring these outcomes, we will test whether participants update their beliefs about the effectiveness of the program by asking them to forecast the results of the full-scale experiment. Again, this forecast will be incentivized in the same way as the previous forecast task. We also ask them their

posterior beliefs about the efficacy of small incentives using a more stringent categorical outcome and their support for experimental evaluations. We then re-elicit respondents' (posterior) institutional trust beliefs. To validate our survey results, we also ask participants to complete a donation allocation task in which they must allocate $30 either to a charity that conducts experiments to measure the impact of its programs or to a similar charity that does not conduct experiments. This decision is incentivized in that 10% of completed surveys are randomly selected to implement their donation preferences and receive a donation receipt for tax purposes.

The survey concludes with the collection of basic demographic data on the participants, such as: education level, age, gender, household income, and political orientation. Finally, we will provide participants with some free learning resources on policy experimentation and scaling up successful interventions to gauge their interest in evidence-based policymaking.

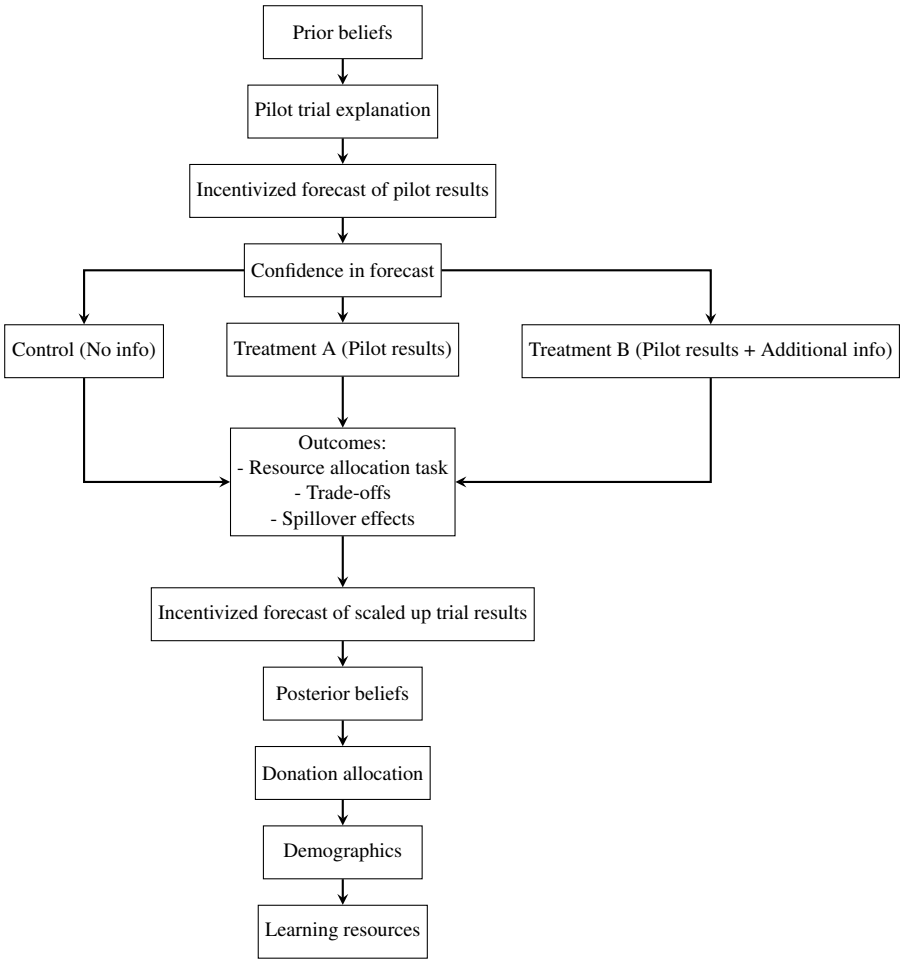

Figure 1

3 Hypotheses and Primary Outcomes

In this pre-registered experiment, we seek to answer the following questions: (i) How does the general public beliefs and preferences compare to those of policymakers?; (ii) Does learning about the (potentially disappointing) results of a study affect citizens' preferences for public spending categories; Does learning about the study results affect (iii) support for experimental evaluations and (iv) trust in institutions; and lastly, (v) Can potential experimentation aversion be overcome by means of a light-touch scalable information provision intervention?

The previously implemented survey experiment on a sample of policymakers showed that policymakers developed an aversion towards experimental evaluation in public policies as a result of learning

the unexpected results of a trial. We hypothesize that the general public will also react to this information, but the direction of the effect on this sample is less clear. Exposure to conflicting information may lead to cognitive dissonance, reducing support for trials, or alternatively, it may increase intellectual humility and the demand for further experimentation. Thus, our first hypothesis is the following:

H1: Treated respondents will exhibit a change in support for experimental evaluations of public policies.

We will examine this hypothesis by analyzing the difference between the control and the treatment group that only sees the trial results (without additional information) on their preferences for public spending allocation and changes in their support for RCTs (see the empirical strategy in Section 4).

It is possible that the general public will react in a similar fashion as the policymakers, thus developing an experimentation aversion upon learning the results of the trial. Our second hypothesis is that, conditional on the first treatment having a negative effect on support for experimental evaluations, a complementary brief educational information treatment can mitigate these effects. Thus, our second hypothesis as as follows:

H2: Exposure to an educational information treatment will affect support for experimental evaluations of public policies, mitigating possible backfire effects from learning the trial results.

In this study we are also interested in providing novel evidence on whether the adoption of a more transparent and agnostic approach to public policy, which involves greater use of experimental evaluations, affects trust in institutions. Trust in institutions is a fundamental pillar of a functioning democracy, and evidence-based policymaking is often seen as a way to enhance this trust by promoting institutional accountability and transparency. However, it is not clear whether the public responds positively or negatively to new information that reveals a program's ineffectiveness. This study seeks to fill this gap by examining changes in trust. Understanding these dynamics can inform how evidence-based practices are communicated and implemented to maintain or enhance institutional trust.

H3: The treatment will affect trust in the public institutions responsible for administering the pro-

gram being evaluated.

The adoption of evidence-based policymaking and transparency in revealing ineffective programs could enhance public trust if citizens appreciate how the scientific process works, but could also diminish if citizens only reward positive trial outcomes. In this regard, we are interested in observing how respondents update their institutional trust beliefs between the two treatments, thus our last hypothesis is the following:

H4: Exposure to an educational information treatment will affect trust in the public institutions responsible for administering the program being evaluated.

4 Empirical strategy

We will evaluate the primary outcomes - namely, support for experimentation and trust in institutions, using the following framework:

$$\Delta Beliefs_{ij} = \beta_0 + \beta_1 Treat_{ij} + \epsilon_{ij} \tag{1}$$

Where $\Delta Beliefs$ is the difference between the posterior and the prior beliefs of each respondent *i* on outcome *j*, which can be either support for experimentation or institutional trust outcomes. The treatment is a categorical variable that can be one of three values: control or one of the two treatments. We estimate this model using OLS, clustering standard errors at the individual level.

To account for different baseline priors, we will report results of the regressions on the same set of primary outcomes using the empirical strategy comparable to Hjort et al. (2021) and the best practice recommended by Haaland et al. (2023):

$$Posterior_{ij} = \beta_1 Prior_{ij} + \beta_2 Treat_{ij} + \epsilon_{ij} \tag{2}$$

We will extend this model by interacting the treatments with the baseline prior beliefs, to examine whether the treatments have heterogeneous effects along the prior beliefs dimension.

$$Posterior_{ij} = \beta_1 Prior_{ij} + \beta_2 Treat_{ij} + \beta_3 Prior_{ij} \times Treat_{ij} + \epsilon_{ij} \tag{3}$$

Controls. We will report the output of our regressions without and with the introduction of a set of covariates of interest, namely: a dummy for whether they heard of 529 before, the average score they give to their agreement with whether 529 plans are useful and equitable, a female dummy, an age group categorical variable, a dummy for whether the respondent pursued any college education, household's income group. In a series of robustness checks, we will also include the number of seconds the respondent spent reading the trial information page, as measured by a hidden timer, as a proxy for attention and understanding of the trial. We will also show additional regressions where we will include dummies for whether the respondent identifies as a Democrat or a Republican, and a dummy for whether their political party aligns with that of the Governor in their state at the time of completing the survey, as a potential source of bias in institutional trust outcomes.

Decision rules for dropping observations. All participants who do not complete the questionnaire will be excluded from the sample. We will also drop (and not compensate) respondents who complete the survey in less than 5 minutes, the minimum amount detected in our pilot.

4.1 Secondary outcomes

We will also implement a series of OLS regressions on a set of secondary outcomes, for both the policymakers' and the general public's samples. These outcomes reflect the typical pitfalls of scaling up interventions (List, 2024). The goal of these additional analyses is twofold: (i) provide novel evidence on how policymakers and the general public reason about scaling up, and whether overoptimism (i.e., a high expected effect of the intervention) biases these perceptions; and (ii) qualitatively compare policymakers' and the general public's preferences and beliefs. The list of secondary outcomes is the following:

- Preferences for public spending allocations across categories (namely: replicating the seeding trial with a different audience sample, conducting the same trial with a larger sample, increasing funding for business-as-usual programs and initiatives, and conducting a new, different trial testing a different intervention)
- Preferences for public spending on scaling up (how to allocate a fixed amount of money between evaluating the effectiveness of a large-scale policy that has been in place for several years and evaluating the effectiveness of scaling up a pilot policy that has shown promising initial results)
- Beliefs of positive and negative spillover effects from interventions aimed at increasing take up

of college savings accounts (namely: other savings crowding out, increase in income inequality, word of mouth, and increased financial literacy)

5 Power Calculations

During the first week of October 2024, we conducted a pilot with N=60 Prolific respondents to obtain a baseline estimate of our outcomes for power calculations. Since we are tracking five main outcomes, we perform power calculations using an adjusted Type I error rate with a Bonferroni correction, dividing α by the total number of comparisons m, thus using a α of 0.001. Using the pilot results, we impute the treatment effects based on equation (1) (using only the treatment where we provide respondents with the pilot results but no additional information, since our expectation is that the second treatment will be more comparable and potentially statistically less distinguishable from control) and the standard deviation of the outcome variable for the full sample. If we run 10,000 simulations on a predetermined sample of N=1,200 respondents, which is dictated by our budget constraints, the statistical power calculations yield the following results:

Table 1: Power calculations on primary outcomes

Δ *Posterior-Prior Beliefs:*	Treat. effect coefficient	Outcome SD for the whole sample	1 - β
Support for experimentation	-0.1980807	1.183643	0.6467
Competence of State Treasurer's Office	-0.3288583	0.8299933	1
Benevolence of State Treasurer's Office	-0.3261771	0.9282328	0.9998
Integrity of State Treasurer's Office	-0.2286839	0.9524026	0.9525

While we may under-powered for the first outcome, we expect to be powered to detect more granular effects using equation (2) estimation strategy.

6 IRB Approval and consent

The proposal has been approved by the University of Chicago Ethics Committee on the 20th of June, 2024 (Approval No. 24-0796).

References

DellaVigna, S., W. Kim, and E. Linos (2022). Bottlenecks for evidence adoption. Technical report, National Bureau of Economic Research.

Garcia-Hombrados, J., M. Jansen, Á. Martínez, B. Özcan, P. Rey-Biel, and A. Roldán-Monés (2024). Ideological alignment and evidence-based policy adoption.

Haaland, I., C. Roth, and J. Wohlfart (2023). Designing information provision experiments. *Journal of economic literature 61*(1), 3–40.

Hjort, J., D. Moreira, G. Rao, and J. F. Santini (2021). How research affects policy: Experimental evidence from 2,150 brazilian municipalities. *American Economic Review 111*(5), 1442–1480.

List, J. A. (2024). Optimally generate policy-based evidence before scaling. *Nature 626*(7999), 491–499.

Mayer, R. (1995). An integrative model of organizational trust. *Academy of Management Review*.

Toma, M. and E. Bell (2024). Understanding and increasing policymakers' sensitivity to program impact. *Journal of Public Economics 234*, 105096.

Vivalt, E. and A. Coville (2023). How do policymakers update their beliefs? *Journal of Development Economics 165*, 103121.