WORKING PAPER · NO. 2024-157

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

Guglielmo Briscese and John A. List

DECEMBER 2024

THE UNIVERSITY OF CHICAGO
BECKER FRIEDMAN INSTITUTE FOR ECONOMICS

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

Toward an Understanding of the Political Economy of Using Field Experiments in Policymaking
Guglielmo Briscese and John A. List
December 2024
JEL No. C9, C93, H4, H41, O12, O36, P1

ABSTRACT

Field experiments provide the clearest window into the true impact of many policies, allowing us to understand what works, what does not, and why. Yet, their widespread use has not been accompanied by a deep understanding of the political economy of their adoption in policy circles. This study begins with a large-scale natural field experiment that demonstrates the ineffectiveness of a widely implemented intervention. We leverage this result to understand how policymakers and a representative sample of the U.S. population update their beliefs of not only the policy itself, but the use of science and the trust they have in government. Policymakers, initially overly optimistic about the program's effectiveness, adjust their views based on evidence but show reduced demand for experimentation, suggesting experiment aversion when results defy expectations. Among the U.S. public, support for policy experiments is high and remains robust despite receiving disappointing results, though trust in the implementing institutions declines, particularly in terms of perceptions of competence and integrity. Providing additional information on the value of learning from unexpected findings partially mitigates this trust loss. These insights, from both the demand and supply side, reveal the complexities of managing policymakers' expectations and underscore the potential returns to educating the public on the value of open-mindedness in policy experimentation.

Guglielmo Briscese
University of Chicago
5807 S Woodlawn Ave
Chicago, IL 60637
gubri@uchicago.edu

John A. List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and Australian National University
and also NBER
jlist@uchicago.edu

1 Introduction

Empirical economics has made important strides over the past half century, and there is perhaps no area that has achieved greater gains in policy circles than field experiments. The roots of this movement can be found in the social experiments of the 20th century (Levitt and List, 2009). While the social experiments have tended to be "black box" in the sense that packages of services and incentives were proffered, more recent field experiments have played an important role in the discovery process by allowing policymakers to make stronger inference on key pieces of a program by splicing large interventions into smaller entities. Similar to the spirit in which astronomy draws on the insights from particle physics and classical mechanics to make sharper insights, field experiments have given the policymaker a toolkit to develop a deeper understanding of the shared behavioral principles to advance optimal policies.

Even though field experimental gains in policy circles are noteworthy, some researchers question why field experimental methodologies are not used to an even greater extent in policymaking (Carattini et al., 2024; Dur et al., 2024; Mazar et al., 2023). We view this "uptake problem" arising from two distinct sources: scalability and uncertainty. First, many experimentally-tested programs fail to deliver their promise at scale (List, 2022). Likewise, policies that have proven effective in small-scale trials are expanded with little verification that their efficacy will sustain at scale, undermining confidence in the science. Second, policymakers are generally in favor of experimentation and are open to implementing solutions that worked elsewhere; however, the unpredictable nature of experimentation with humans means that results are oftentimes surprising, yielding insights that may challenge a person's preconceived beliefs about a policy's efficacy. This uncertainty might, in turn, impose unforeseen costs on the policymaker. While there is a burgeoning literature exploring the science of scaling explanation (Al-Ubaydli et al., 2017; Mobarak, 2022; Agostinelli et al., 2023; Wang and Yang, 2021; Larroucau et al., 2024), there is little systematic evidence on the effects of policy uncertainty. That is, despite the acknowledged importance of maintaining an agnostic mindset when implementing trials, little is known about how policymakers and the general public react when learning about trial results that do not meet their expectations.

This study contributes to our understanding of the political economy of field experiments in policymaking by using a two-step field experimental approach. First, we conduct a large-scale natural field experiment to evaluate the efficacy of small financial incentives to increase participation in college savings accounts — an intervention that has been implemented in various U.S. states but has not been previously evaluated experimentally. These incentives are often used across several other public programs with the premise that small financial rewards can increase program attractiveness and help citizens outweigh the burden of sign up costs. Yet their impact has remained largely speculative. We find evidence of a reasonably tight null result: our incentives were not effective in enhancing participation in college savings accounts.

Our second step is to explore the political economy repercussions of this result. To

do so, we examine how policymakers respond to novel, potentially unexpected findings, by conducting a survey experiment with a sample of U.S. state policymakers responsible for administering college savings accounts and similar state-run savings products. The survey incorporated an incentive-compatible mechanism to elicit policymakers' forecasts of the trial outcomes before randomly revealing the actual results. We then replicated this policymaker survey experiment using a representative sample of Americans to explore how the voting constituency responds to such information.

Our findings reveal five facts on the political economy of using field experiments for policymaking. First, policymakers update their opinions based on new scientific findings. After learning the true trial results, treated policymakers revised their beliefs and shifted their stated resource allocation preferences, indicating a reduced focus on scaling the intervention and increased interest in funding new evaluations instead of maintaining the status quo. Second, policymakers update too broadly, spreading received learnings from the narrow intervention to their beliefs about the general scientific approach. This finding reveals the difficulties of human experimentation for policymakers: given the inherent uncertainty in results, policymakers might become more pessimistic about the entire scientific enterprise after receiving one set of results that does not conform to their priors. Third, striking similarities exist between a representative sample of US citizens and policymakers. For example, similar to policymakers, the public is (even more) optimistic about the potential for financial incentives to improve program participation rates and upon receiving the experimental results had remarkably similar stated preferences for how public resources should be re-allocated between scaling and further evaluations.

Fourth, while similarities exist, there are striking dissimilarities between a representative sample of US citizens and policymakers. For example, at odds with policymakers, public support for policy experiments is high and remains robust even after learning of the disappointing trial results. However, trust in the government institutions responsible for implementing the trial declines, particularly with regards to perceptions of competence and integrity. Our fifth, and final fact is that citizen trust of policymakers can be partly restored with education. We find that a randomized treatment that provides respondents with additional information explaining the value of policy experiments, emphasizing the importance of learning from unexpected results, partially mitigates the loss of trust revealed in our fourth fact. Such campaigns have the potential to improve significantly the perceptions of competence and increasing perceptions of benevolence. However, trust in the integrity of the government institutions remains more vulnerable to erosion, underscoring the challenges of maintaining public trust when trial results do not align with expectations.

These five facts paint a complex picture of the political economy of policy experimentation while also highlighting a significant challenge to advancing evidence-based policymaking at scale. While support for policy experiments is high among both citizens and policymakers, disappointing trial outcomes can trigger experimentation aversion among policymakers and erode institutional trust among the public. Experimentation inherently involves uncertainty since there is always a chance that results will reveal a policy's ineffectiveness. Previous survey experiments studying policymakers' and politicians' belief-updating have typically

focused on "good news" scenarios, where trials showed policy effectiveness (Hjort et al., 2021; Garcia-Hombrados et al., 2024), overlooking the full range of possible outcomes. Successful policy experimentation requires an agnostic approach and open-mindedness from both citizens and policymakers for repeated implementation at scale.

Our results further underscore the importance of transparent communication and education about the role of policy experiments, especially in managing expectations around the unpredictability of outcomes. Additionally, our findings reveal that policymakers' overoptimism, followed by aversion to experimentation after negative results, poses a key barrier to sustaining experimentation efforts. This challenge might be mitigated by fostering a culture of learning from all outcomes — positive or negative — and aligning expectations more effectively. Below, we discuss potential actions that governments might consider if they wished to promote evidence-based policy-making in a way that sustains public trust while encouraging policymakers to remain committed to experimentation, even when results are unexpected or disappointing.

We view our findings as important because they suggest that for field experiments to grow in import, both policymakers and the public should be better informed and more open to the iterative process of scientific experimentation. This movement can lead to more robust and credible policy decisions, ultimately fostering greater trust in science and government. In short, our results highlight the importance of improved communication and education around the role and value of field experiments in policymaking.

Our study speaks to several strands in the literature. First, it expands the emerging, albeit still limited, literature that specifically examines policymakers' demand for evidence-based policies. Surveying a sample of policy practitioners from the World Bank and the Inter-American Development Bank, Vivalt and Coville (2023) find that policymakers are optimistic about the effectiveness of development programs and that they update their beliefs when presented with robust evidence. Importantly, they find that policymakers update more on 'good news'. Our study shows that optimism about program effectiveness is equally common among U.S. state policymakers, but the general public is even more optimistic in comparison. We find that policymakers update beliefs also when presented with 'bad news', but provide novel insights on the implications that this belief-updating has for resource allocation preferences and demand for policy experimentation.

Relatedly, using a sample of U.S. federal government employees, Toma and Bell (2024) show that policymakers find it difficult to translate evaluation studies conducted in similar policy areas into actionable changes to their policies and programs, and that making the results of previous evaluation studies more easily interpretable changes policymakers' sensitivity to policy impact. In our study we show policymakers the results of a randomized control trial (RCT) that tested a policy they are familiar with and that many states have been implementing for several years. We also use a simple and incentive compatible elicitation of forecast effects, which we show replicates well on a sample of the general public. Ours is also the first study that uses a survey experiment on a sample of U.S. state policymakers who are directly responsible for managing large budgets and overseeing programs affecting

millions of citizens.

Another small number of studies focus on understanding politicians' - instead of policymakers - demand for evidence-based policies. Hjort et al. (2021) shows that politicians are willing to update their beliefs when presented with new evidence, and that they are eager to implement ideas that proved effective elsewhere. Garcia-Hombrados et al. (2024) show that this tendency is stronger when research findings come from an ideologically aligned source. These studies importantly show the promise of sharing research findings to political decision-makers to spread adoption of policies that worked elsewhere. Yet, less is known about how these key decision-makers are willing to divert resources away from and the trade-offs they consider when deciding which programs to evaluate or scale when presented with new evidence. Moreover, this line of work has not explored how policymakers update their beliefs and change their resource allocation preferences when evaluations yield unexpected disappointing results, especially if they relate to a program that is already being implemented.

A second area of research our five facts contribute to relates to the interplay of preferences between evaluations and scaling among policymakers and citizens, and how they view such trade-offs. In this manner, our results speak to how people account for basic "vital signs" of scaling (List, 2024b). For example, our finding that policymakers revise their beliefs and resource allocations after new scientific results highlights the importance of avoiding false positives. Ensuring that positive results are genuine and not just random chance is crucial for maintaining policymakers' trust in field experiments. Notably, while policymakers adjust their resource allocation preferences based on new evidence, their willingness to engage in future experiments can be adversely affected by disappointing results. This underscores a paradox in scaling scientific interventions, where the inherent uncertainties that make science valuable can also foster skepticism. In this regard, we demonstrate that learning from trial evaluations can encourage more critical thinking about the efficacy of policies at scale. These insights are crucial for designing strategies that ensure sustained support for experimental methods at scale, ultimately driving more informed and effective policy decisions.

Finally, we contribute to the literature on institutional trust, which has traditionally been a subject of research in public administration and management science but has not been applied to this context. While one might expect that a government committed to evidence-based policymaking signals transparency and efficient use of taxpayers' money to its constituents, our findings suggest this may not hold true when evaluation outcomes contradict citizens' prior beliefs. Like other studies on non-U.S. populations (Dur et al., 2024), we find that Americans support the use of experiments in public policy, even when results don't align with their expectations. However, we provide novel evidence that unexpected trial results can undermine trust. While educational interventions can mitigate these effects, more research should examine the relationship between policy experimentation and trust in government.

The remaining parts of our study are structured as follows. In the next section, we provide background information about the policy context of our study. Section 3 summarizes the pre-registered design and results of our large-scale natural field experiment. Section 4 reports the design and results of the survey experiments on the sample of policymakers and

the general public. The last section concludes the paper.

2 Institutional Background of 529 Plans

College Savings Accounts, commonly known as 529 plans, are state-administered programs designed to help families save for their children's future college expenses, potentially reducing the need for student loans. These accounts have minimal impact on financial aid eligibility, as only 5.64% of the total assets in a 529 plan are considered parental assets. Named after Section 529 of the IRS Code, these plans allow contributors to either prepay a beneficiary's qualified higher education costs or contribute to an account dedicated to covering future postsecondary expenses. In 2001, the Economic Growth and Tax Relief Reconciliation Act made earnings on these plans completely tax-free (with this provision becoming permanent in 2006). Today, all 49 states and the District of Columbia offer 529 plans, many of which also provide state income tax deductions for contributions. As of June 2024, there were 16.25 million active 529 accounts that totaled $450.5 billion.

Each state is responsible for administering its 529 plans(s)[1], which includes promoting the program to families within the state and nationwide. These efforts often involve targeted outreach initiatives to educate parents, as well as broader marketing campaigns to raise awareness about the program and its benefits. To boost participation, several states offer small financial incentives, sometimes targeted at underrepresented communities. These incentives typically take the form of seeding (initial deposits) or matching contributions. Such incentives are intended to attract parents' attention, reduce perceived administrative barriers to opening an account, and encourage regular contributions by demonstrating the benefits of compound interest.

As part of these efforts, every year many states offer various time-bound financial incentives to encourage families to open 529 accounts, sometimes around May 29th (5/29) as part of the '529 day' which many states officially recognize as 'College Savings Day'. For example, in 2024, California offered a $50 match for families who open a new ScholarShare 529 account between May 20 and May 31, 2024 with $50 or more. Florida residents that enrolled in a new Florida 529 Savings Plan by June 23, 2024, received a $50 account contribution to kick-start their savings journey. Utah residents were eligible for up to a $40 match if they opened a new account during the month of May [2]

[1] A range of public bodies administer 529 plans in different states – some are rooted in educational authorities (e.g., AK, MA, UT), some are quasi-independent (e.g., VA), many are Treasurer's Offices (e.g., CT, IL, CA), and others are the Comptroller (TX).

[2] More information about similar initiatives ran by each state can be found online here. Other initiatives also exist across multiple states that are not necessarily time-bound and are supported by more intensive marketing programs. For example, the Illinois First Steps program provides a one-time $50 seed deposit for state residents when they open an account for a child born or adopted on January 1, 2023 or later. Other examples include Oregon's $25 "Baby Grad" program and Rhode Island's $100 "CollegeBound Starter", and Maine's $100 seed deposit. The results of our trial are thus informative also for more open-ended and more targeted outreach initiatives.

While these initiatives share the common goal of encouraging parents to open 529 accounts early, ideally when children are young to maximize long-term savings, the effectiveness of these financial incentives in driving account uptake has yet to be rigorously established. This is our first goal of this study: provide empirical evidence from a natural field experiments testing the efficacy of incentives.

3 Natural Field Experiment

We partnered with the Illinois State Treasurer's Office (ILSTO) to implement a large-scale natural field experiment to test the efficacy of a scalable intervention consisting of offering small financial incentives to increase uptake of 529 plans in Illinois. There is a dearth of evidence on effective strategies to encourage greater participation in 529 plans, and the evidence on the efficacy of small financial incentives to increase savings is mixed, leaving us with the need to implement a robust evaluation to generate novel causal evidence.[3]

To generate the data, we partnered with ILSTO to implement a field experiment testing the efficacy of financial incentives in increasing 529 plan sign-ups and contribution rates. The aim was to generate novel evidence on scalable interventions that the state government could deploy. To ensure scalability, we focused on incentive amounts that were feasible within budget constraints in the event of a high state-wide take-up rate and leveraged the existing communication platform to minimize the risk of scaling bottlenecks (DellaVigna et al., 2024).

For the experiment, we used email addresses of Illinois residents who had previously interacted with state government agencies. We limited the sample to those with a valid email address, residing within Illinois, and not already associated with a 529 account, yielding a final sample of 734,070 email recipients. Using Census data, we matched recipients' zip codes to ensure coverage across the state. Every zip code in Illinois had at least one recipient, with user density ranging from 1 to 31 users per 100 residents (96% of zip codes had between 2 and 9 users per 100 residents). Overall, the database represented 6% of non-Chicago residents and 5% of Chicago residents.

Although the resulting dataset lacked detailed household information, such as the number of children, it was ideally suited for testing a low-cost, large-scale state-wide marketing campaign intervention. To ensure that we were actually conducting a natural field experiment, our approach closely mimicked the typical capacity and delivery methods

[3]Mason et al. (2010) utilized data from a Child Development Account initiative to examine the relationship between financial incentives and saving outcomes in programs specifically for low- and middle-income children. Their analyses suggest that varying types of financial incentives had different associations across savings outcomes, and further flagged the possibly regressive nature of the plans. In a pilot study ran in partnership with the pediatrician's offices of the Boston Medical Center, new parents were helped by the birth registry staff to open a 529 account, thus not providing financial incentives but reducing the psychological costs of administrative burden (Herd and Moynihan, 2019); the study found that only 6% of parents ended up opening an account (Tummala et al., 2022).

used by our partner agency and other state government agencies. To further minimize additional costs, and maintain naturalness and scalability, we utilized the agency's existing email marketing platform and adapted their current electronic marketing materials for the experiment's content.

From the contact list, we randomly selected 7,500 individuals for a soft launch, followed by a larger sample of 150,000, also randomly drawn from the remaining subject pool. Depending on the initial take-up rates and budget availability, the ultimate goal was to contact all recipients in the database. Both the soft launch and larger sample were representative of the broader pool and balanced across median income levels at the zip code level. All recipients received similar emails, differing only by one sentence, which offered a seeding incentive. Email recipients in both samples were randomized into one of four conditions: no incentive, or a seeding incentive of $10, $50, or $100. Treated individuals were informed that the money would be deposited into their new 529 accounts if they signed up within 30 days using the same email address at which they received the marketing email. The digital content used in the trial is shown in Section 7.1.

3.1 Field Experimental Results

In October 2023, we implemented the soft launch following the pre-registration of our analyses (AEARCTR-0012055). The results showed a relatively high email open rate (40%) across both control and treatment groups, but only four recipients opened an account: one in the control group, two in the $50 incentive group, and one in the $100 incentive group. Based on these results, we revised the power calculations in our pre-analysis plan and proceeded with the larger trial to ensure that the pilot did not miss out on potential new account holders due to a lack of important observable characteristics, such as whether the email recipient was a parent[4]. The full-scale trial, involving 150,000 recipients, revealed that over half of the recipients across all groups opened the email, and approximately 2% clicked on the provided link to learn more about 529 plans. These engagement metrics are notable as they are substantially higher than the average engagement rate records in other email marketing experiments (Sahni et al., 2018). This further supports that there is merit in exploring the efficacy of this outreach campaign to increase program uptake. We found no significant differences in these engagement metrics across the randomized groups (see Section 7.1 for a visual summary). Over the four-week observation period, no one in the control group opened an account, compared to one person in the $10 seed group, 11 in the $50 group, and 9 in the $100 group.

These experimental results, which share similarities to some charitable giving research using mailers that attract little support (see, for example, Landry et al. (2010)), suggest that non-targeted, large-scale marketing campaigns, even when supplemented with

[4]Note that, although parents may have a stronger interest in the program, anyone can open a 529 account. Thus, our large pool would have likely included parents as well as relatives (e.g., grandparents) or friends who would also be interested in opening an account for a child.

small financial incentives, are unlikely to be very effective in increasing participation in 529 plans. Our engagement metrics allow us to rule out lack of compliance as the key mechanism, and offer a reliable finding that the treatment itself is not very impactful. Since many U.S. states adopt similar initiatives offering time-bound financial incentives to encourage uptake, these findings could have significant implications for national policymaking. If policymakers take these results into account, then they may influence future resource allocations and adjustments in program design.

Table 1 provides some insights into how a policymaker can make use of our null results. The table provides the post-study probability (PSP) that a real relationship exists, given that we find a null result, for different power levels (β) and prior probability that the relationship exists (π). The table shows that for well powered studies (0.80), the impact of a null result on the PSP is higher for relationships with high priors than for relationships with low priors. Given that there are high priors about the existence of a real relationship between seeding incentives and take up rates of 529 accounts, we might assume that π is 0.90. In this case, with our level of experimental power (0.80), our null result has a sizable impact on beliefs: it moves the beliefs from 0.90 to 0.65 that a real relationship exists between incentives and take-up. If, instead, we assume that π is 0.80, then our post study probability is 0.46. In addition, Table 1 reveals that if a well-powered replication of our study yields another null result, the post study probability decreases even further (in the case of our study moving π from 0.80 to 0.46, insights from the replication study would move the PSP even lower, to roughly 0.15). While many in the profession view null results as contentless, this exercise highlights the inferential import of well-powered null results. In fact, in certain cases null results are more informative than rejections of nulls (List, 2024a).

Table 1: Post Study Probabilities after a null result for different levels of power and prior probability that a real relationship exists.

	Power				
π	**0.20**	**0.30**	**0.50**	**0.70**	**0.80**
0.10	0.09	0.08	0.06	0.03	0.02
0.20	0.17	0.16	0.12	0.07	0.05
0.30	0.27	0.24	0.18	0.12	0.08
0.40	0.36	0.33	0.26	0.17	0.12
0.50	0.46	0.42	0.34	0.24	0.17
0.60	0.56	0.53	0.44	0.32	0.24
0.70	0.66	0.63	0.55	0.42	0.33
0.80	0.77	0.75	0.68	0.56	0.46
0.90	0.88	0.87	0.83	0.74	0.65

4 Survey Experiments

Policy experiments are an essential element of evidence-based policymaking, ensuring that taxpayer funds are directed toward programs that have been shown to work. By testing and evaluating policies, these experiments provide valuable insights that help governments make informed decisions about scaling or discontinuing programs. However, the viability of this approach depends on both the demand and supply of policy experimentation. Policymakers must be willing to adopt experiments, even when the results may reveal that their programs are ineffective, and the public must recognize and support the use of such experiments, even when outcomes may contradict their expectations or ideological views. Without these conditions, there is a risk that the "file drawer" problem — where only studies with favorable results are published in academic journals — could extend to public policy, with potentially harmful consequences for millions of people.

To better understand these dynamics, we conducted two survey experiments: one with policymakers working at state agencies overseeing 529 plans and similar savings programs, and another with a representative sample of Americans. These two surveys explore how both groups perceive the value of policy experiments and their willingness to embrace evidence-based approaches in light of received results.

4.1 Policymakers' survey

We recruit a sample of policymakers for whom our field experimental results represent valuable knowledge that can influence their policy decision-making. To do so, we leverage the Management Training Symposium annual national conference hosted by the National Association of State Treasurers (NAST), which serves as a regular opportunity for treasurer's office staff across the country to share knowledge and best practices.[5] Conference attendees include staff responsible for administering 529 plans in their state as well as staff working on the promotion and management of similar financial products that often share similar goals and strategies.[6] The NAST annual training conference represents an ideal setting to reach a pool of public servants who are familiar with the 529 program and would deem the findings

[5]The National Association of State Treasurers (NAST) is the collective authority on practices and policies related to finance and investment for state government. It brings together state treasurers and state finance officials with comparable responsibilities from the United States, its commonwealths, territories, and the District of Columbia, along with employees of these agencies. Among NAST's goals is knowledge-sharing of best practices, offering educational conferences and webinars, a variety of working groups, policy advocacy and publications that provide information about developments in public finance. NAST's mission includes holding members and staff "accountable and act with integrity and transparency representing the best practices of their office and the Association", and promoting and facilitating "the open and bipartisan exchange of time-honored and innovative ideas and information, and seek out external collaborations wherever appropriate, to provide objective analyses and viable solutions to address state financial and investment concerns".

[6]The savings programs administered by treasurer's offices include national initiatives such as the 529 College Savings Accounts, the Achieving a Better Life Experience Program (ABLE), and retirement savings programs for workers who do not have access to an employer-sponsored retirement plan.

from our field experiment relevant.

We partnered with the organizers of the NAST Treasury Management Training Symposium to send an email invitation to all registered attendees — both in-person and virtual — for the 2024 conference held in Pittsburgh, PA, on May 20-23, 2024. The email, sent on May 12, targeted approximately 1,200 individuals. It was intentionally brief and included a link to an online survey. To encourage participation, we gave the respondents an opportunity to earn up to two $25 Amazon vouchers, based on their survey responses. We chose this incentive structure, offering larger prizes rather than a smaller flat fee, due to the unique nature of our participant pool and recruitment process. By the time we closed the survey, the night before the conference began, we had collected 143 responses. Some respondents didn't complete the final module, but nearly completed the entire survey, hence our sample size changes slightly on some outcomes. We do not notice any significant predictor and no differences between the control and treatment group on survey drop out decisions.

Two important points are worth noting about our sample size. First, treasurer's offices, like the Illinois office involved in our field experiment, tend to be relatively agile public sector agencies with small staffs that manage large portfolios of programs affecting millions of people in their states. For instance, the Illinois office has only three full-time employees (FTEs) exclusively dedicated to 529 plans, three FTEs working on ABLE accounts, and three FTEs focused on retirement savings [7]. Second, recruiting policymakers is notoriously challenging, and similar studies often achieve comparable sample sizes.

For example, Vivalt and Coville (2023) obtained a final sample of 378 policymakers by partnering with two of the world's largest international organizations — the World Bank and the Inter-American Development Bank — and conducting recruitment at multiple events in countries such as Portugal, Mexico, Nigeria, and Senegal over nearly two years. Their recruitment efforts even included setting up a table by the cafeteria at the organizations' Washington D.C. offices to engage passersby. Similarly, Toma and Bell (2024) recruited 191 employees from 22 out of 24 U.S. federal agencies over six months through word of mouth and snowball sampling, with support from the U.S. Office of Evaluation Sciences.

Other studies that recruited politicians instead of policymakers have faced similar challenges. Hjort et al. (2021) recruited participants through two national conventions of Brazilian mayors, a national conference, and 12 regional conferences held across Brazilian states over one year, achieving sample sizes of 486 and 702 participants. These experiments involved half-hour self-administered sessions using tablets, with research assistants actively recruiting participants during conference breaks. In a forecasting survey that more closely resembles our design, Jilke et al. (2024) contacted 815 county heads and mayors or municipal government representatives of towns with 30,000 or more residents in Germany, achieving a a final sample of 88 total responses. Lastly, Dur et al. (2024) invited all 725 members of Dutch national and regional parliaments and obtained a sample of 126 complete responses.

[7] Additional employees dedicate a portion of their time to support these program areas, such as the legal department.

These noteworthy studies highlight the challenges of conducting experiments with policymakers, as recruitment is often both costly and time-consuming. In contrast, our study was conducted with a zero-dollar research budget over a one-week period. Additionally, our recruitment approach — which is similar to Toma and Bell (2024) — attempted to minimize the risk of response contamination by recruiting participants before the conference began. This approach reduces the chance that subjects discuss the study with each other, which potentially influences responses.

Respondents took an average of 18 minutes and a median of 14 minutes to complete the survey, which consisted of 13 questions including a demographics module. The survey was anonymous, and to maintain confidentiality, we did not ask for the name of their employer or their state of residence[8]

4.1.1 Survey design

Survey invitees were first prompted to read the Participant Information Statement before agreeing to participate (see survey flow in Figure B.6). The survey began with a brief explanation of the pilot experiment, which tested the effectiveness of small incentives in increasing 529 plan take-up rates. This included an image of the email used in the trial. To avoid biasing respondents' forecasts, we omitted any mention of the trial partner's name (see Figure B.7 for materials used). Next, respondents were asked to forecast the exact number of individuals who opened an account in each of the control and treatment groups, with the reminder that each group had 1,875 email recipients. By eliciting exact numbers rather than percentages, we aimed to minimize the risk of respondents misinterpreting percentage differences versus percentage point differences. Forecast elicitation was financially incentivized, following the approach of DellaVigna and Pope (2018). Respondents were informed that if their forecast fell within ±30% of the actual results, they would be entered into a drawing for a $25 Amazon gift card. On the same page, we also asked respondents to indicate how confident they were in their forecasts.

Participants were then randomized into a treatment or control group: the control group only saw a thank you message for providing their best guess, while the treatment group saw the results of the pilot experiment. This treatment is meant therefore to learn about the efficacy of a marketing email by seeing the baseline take-up from the control group as well as learn the efficacy of different financial incentives amounts.

After randomization, respondents completed a resource allocation task in which they allocated a hypothetical sum of $100,000 — a typical project budget in state governments — across four initiatives presented in random order. These initiatives included: replicating the seeding pilot trial with a different group of subjects, conducting the same trial with a larger

[8]Ninety out of 143 respondents (63%) chose to share their email addresses, from which we can infer that at least 28 states were represented. Only one respondent worked for Illinois State Treasurer's Office, although from another non-529 unit and is highly unlikely this person knew about the trial prior to completing the survey.

sample,[9] increasing funding for business-as-usual (BAU) programs, and launching a new trial testing a different intervention. The purpose of this task was twofold: first, to assess whether learning the results of an RCT immediately influenced policymakers' preferences for resource allocation towards scaling, replication, experimentation, or non-experimental BAU activities; and second, to provide new evidence on how policymakers navigate trade-offs when deciding which policies to scale or evaluate — an aspect often overlooked in previous studies that fail to capture how policymakers shift resources to support evidence-based policymaking.

Our next survey query aimed to capture the trade-offs policymakers face when deciding which projects to evaluate and which to scale. Respondents were asked to allocate again a hypothetical $100,000 budget between two evaluation projects, presented in random order. The options were: (i) *Robustly evaluate the effectiveness of a large-scale policy that has been in place for several years. There's a risk you might discover it's not working as well as hoped*; (ii) *Robustly evaluate the effectiveness of scaling up a pilot that showed promising initial results. However, there's a risk that these positive effects won't hold true when implemented on a larger scale.* The wording explicitly highlighted these trade-offs, emphasizing the potential risks and uncertainties associated with each option.

The fourth survey question sought to gauge policymakers' beliefs about potential spillover effects when scaling interventions. Respondents were asked to rate, on a 10-point Likert scale, the likelihood that various spillovers would occur when expanding an intervention designed to increase participation in any savings program. This question aimed to provide novel insights into whether policymakers' overoptimism plays a role in determining which programs get scaled, as an expectation of large effects may lead to an underestimation of the challenges of scaling. Additionally, it explored whether learning from past trials (i.e., our treatment effect) could help mitigate this tendency.

After measuring these outcomes, we asked participants to update their beliefs about the effectiveness of the seeding incentives by forecasting the results of the full-scale experiment. Respondents were informed that the trial implementation partner had scaled the experiment to a sample of 150,000 email recipients. Similar to the previous forecast task, they were required to predict the number of accounts opened in the control and treatment groups. This elicitation task was also incentivized, with correct responses within a $\pm 30\%$ range automatically entered into a random drawing for another $25 Amazon gift card. To avoid excessive priming and social desirability bias in the subsequent questions, we did not reveal any of the full-scale trial to respondents, and instead told them that the results would be shared during a session at the upcoming conference.

After the second forecasting task, respondents proceeded to two final modules with questions about their organization and demographics. The organizational module included questions on whether their organization had conducted an RCT or quasi-experimental evaluation in the past five years, if they had used seeding incentives, their operational area (529, ABLE, retirement savings, or other), and their organization type. Within this module, we also included a follow-up to elicit respondents' posterior beliefs about the efficacy of

[9] At this stage of the survey, respondents did not know that a scaled up trial was in fact implemented.

seeding incentives, as per Haaland et al. (2023). Specifically, we asked respondents to rate their belief in the effectiveness of seeding incentives on a 5-point Likert scale ranging from "Highly ineffective" to "Highly effective," with an option for "Don't know." The purpose of this task was to determine whether the treated respondents had internalized the pilot results.[10]

The final demographic module included questions on work experience, expected years remaining in their current job, education, age, and gender. The concluding "thank you" page provided links to two online resources for those interested in learning more about RCTs in public policy: MIT's JPAL non-technical introduction to randomized evaluations and a University of Chicago podcast on scaling promising social programs. These links were displayed in random order, and a hidden click tracker was embedded to monitor engagement.

4.2 General public's survey

To complement our policymaking survey, in October 2024, we conducted a three-arm survey experiment on a representative sample of 1,200 Americans, recruited via the Prolific platform, with the analysis pre-registered (AEARCTR-0014575).[11] The survey structure and questions mirrored those used in the policymakers' survey, with a few key differences. The primary distinction between the general public survey and the previously administered policymakers' survey is the inclusion of questions designed to elicit (a) respondents' support for policy experiments and (b) their trust in public institutions. These questions were asked both at the beginning and the end of the survey, following standard survey experimental practices that measure shifts in beliefs and policy preferences (Haaland et al., 2023; Hjort et al., 2021).

The survey opened by assessing respondents' prior trust in public institutions — specifically, their state treasurer's office, the agency responsible for administering the program they would later learn about, and their state governor's office. We included both government agencies for two reasons: first, to capture potential differences in trust between institutions that respondents may perceive as more or less politicized, with the treasurer's office likely being less familiar to them; and second, to evaluate potential spillover effects in institutional trust. To measure trust in these institutions, we employed a validated methodology widely used in management science and public administration research but less common in economics. This approach, originally proposed by Mayer (1995), breaks down trust perceptions into three dimensions: ability (the institution's competence in delivering

[10]Additionally, this module asked respondents to indicate their interest in signing up for a meeting at the conference or afterward to discuss establishing a national research consortium, which would involve collaborating with external academic researchers on RCTs similar to the one they had just learned about. Respondents could express this interest using a 0-100 probability scale, following the elicitation methodology of Wiswall and Zafar (2015).

[11]We did not pre-register the analysis plan for the policymakers' survey experiment due to an oversight in the days leading up to the conference. However, for the general public's survey, we pre-registered a detailed analysis plan, and we applied the same procedures, decision rules, and analytical methods to both surveys. This alignment strengthens the reliability of our findings from the initial experiment by ensuring they were analyzed in accordance with a pre-specified plan.

quality services), benevolence (the alignment of its motives and values with those of the public), and integrity (its commitment to transparency and honesty).

Respondents were then asked to read a short paragraph explaining experimental methodologies and how they can benefit public policy improvements before expressing their support for this evaluation approach. This module helps control for individual differences in prior knowledge of RCTs and reduces potential experimenter demand effects by ensuring all participants receive the same basic information about the technical aspects and importance of these methodologies in public policy.

The third module gauged respondents' prior knowledge, experience, and beliefs about 529 plans. Regardless of their initial responses, all participants were shown a brief explanation of what 529 plans are and how they can help families save for college, ensuring that everyone had the same baseline understanding. The aim was to control for varying levels of prior knowledge that could influence subsequent responses.

At this point in the survey, the modules mirrored those in the policymakers' survey: respondents read a brief overview of the pilot experiment testing the efficacy of small incentives to increase 529 plan enrollment. They were then asked to predict the pilot's results in an incentive-compatible manner, using the same reward structure as in the policymakers' survey. As in the earlier survey, information about the trial partner's identity was omitted to avoid biasing responses, and participants were asked to imagine the trial took place in their own state. To examine whether respondents took this hypothetical framing seriously, we randomized Illinois participants into two groups: one group received the same neutral wording as the rest of the respondents, while the other group was explicitly shown the name of the trial partner agency.

Participants were then randomized into one of three groups: a control group that only received a thank-you message for providing their best guess; a treatment group that was shown the results of the pilot experiment; and a third group, which differed from the policymakers' survey, that viewed the pilot results alongside a brief explanation of why these results are crucial for informing evidence-based policymaking. This additional trial arm aims to assess whether a light-touch information provision intervention can help mitigate any negative effects that may arise from presenting respondents with the trial results.

In line with the policymakers' survey, all respondents then completed the resource allocation, trade-off, and spillover questions, followed by the incentivized forecast task to guess the effects on the full-scale trial. We also asked for posterior beliefs about the efficacy of small incentives, similar to our approach with policymakers, and their posterior beliefs on support for experimental evaluations in public policy and institutional trust beliefs. To validate our survey outcome on support for RCTs, we asked participants to complete a donation allocation task in which they allocated \$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 was incentivized in that 10% of completed surveys were randomly selected to implement their donation preferences and received a donation receipt for tax purposes.

The survey concluded with the collection of basic demographic data on the participants, such as: education level, age, gender, household income, and political orientation. Finally, we provided participants with the same free learning resources on policy experimentation and scaling up to gauge their interest in evidence-based policymaking. Participants received $2.50 per completed response, which took on average 16 minutes and median 14 minutes.

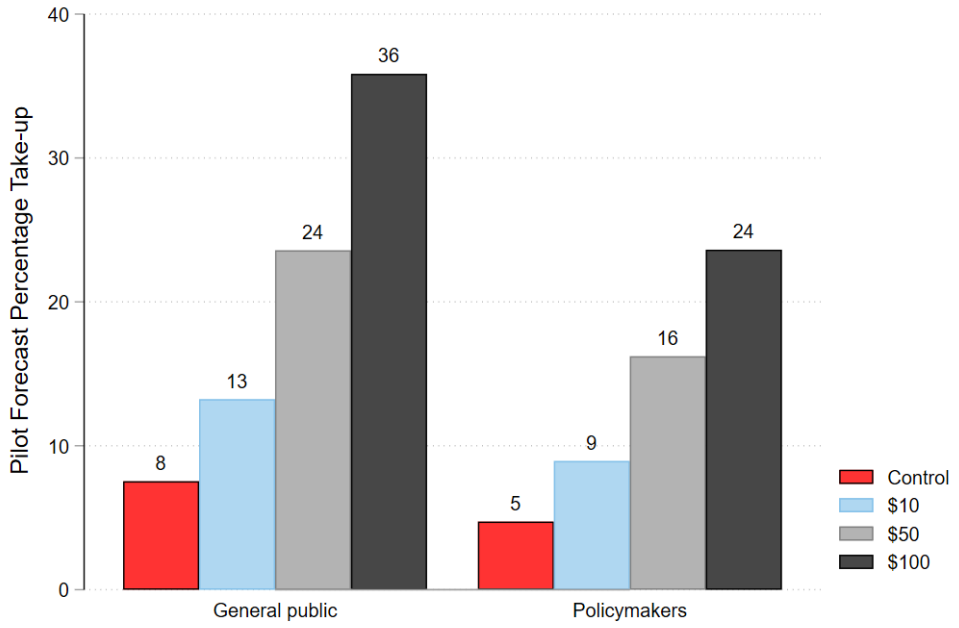
4.3 Survey Results

We present our results in the order of the survey flow, showing side by side the results of policymakers and the general public respondents. We then proceed to discuss the results of the general public's survey experiment on the institutional trust outcomes.

4.3.1 Forecasting and belief updating

Figure 1 summarizes the forecast take-up rates of 529 plans across the four randomized conditions— the control ($0) and the $10, $50, and $100 seeding incentive conditions — among both policymakers and the general public. Several interesting insights emerge. First, both groups are overly optimistic about the 529 plan take-up rates. Policymakers expected 5% of the 1,875 email recipients in the control group to open an account, while the general public predicted a higher rate of 8%. Second, both groups forecast a non-linear increase in take-up rates as the seeding incentive amount rises. Third, the general public consistently overestimates the take-up rate across all conditions compared to policymakers, suggesting that policymakers' expertise may temper overoptimism.

Figure 1: Forecast percent take up across control and treatments, by sample of respondents



Notes: The figure shows the (financially incentivized) forecast percent take-up across randomized conditions - control (no incentive) vs one of three different amounts of seeding incentive, by sample of respondents. General public's N=1,200; Policymakers' N=143

We examine the key correlates of the expected efficacy of small financial incentives by calculating, for each respondent in both samples, the difference in take-up percentage between each treatment condition and the control. To standardize this across individuals, we use the highest value among the three treatment conditions as the most optimistic expected treatment effect. This approach accounts for individual-specific idiosyncrasies in their control group expectations and provides a more comparable measure of anticipated treatment effects.

Among the sample of policymakers (see Table B.4), we do not find any individual characteristic to be significantly correlated with the expected efficacy of financial incentives. While working for an employer that uses similar financial incentives to boost program participation is positively correlated with expected treatment effects, and working specifically on 529 plans is negatively correlated, neither relationship is statistically significant. This suggests that overoptimism in policymakers' treatment effect forecasts may be widespread and not systematically linked to observable characteristics.

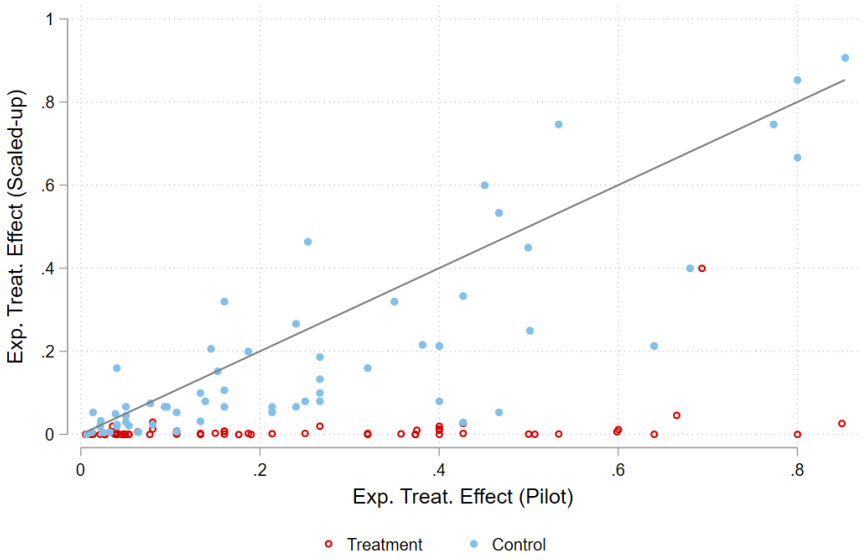
In contrast, the same regression on the representative sample of Americans (see Table B.4) reveals that optimism about treatment efficacy is strongly correlated with respondents' confidence in their forecasts and their belief that the program generates substantial societal benefits. This result provides suggestive evidence that confirmation bias may play a role in explaining individual overoptimism. Alternatively, stronger familiarity with the program — whether respondents were aware of 529 plans or owned a 529 account themselves — shows a

negative correlation with optimism, aligning with the trends observed in the policymakers' survey. This result suggests that greater familiarity with the program provides respondents with more contextual knowledge, leading to forecasts that are closer to the actual take-up rate.

Posterior beliefs. A next empirical question revolves around whether respondents update their beliefs in the expected program take up rate upon receiving this news. We gain insights in this domain by comparing the difference in forecasts between the pilot and the full-scale trial. In the policymakers' sample, half of the respondents saw the results of the pilot prior to reporting their full-scale trial forecast, and in the general public's sample two-thirds - i.e. the two treatment groups - saw the pilot results. It is worth noting that also the forecast for the results of the full-scale trial were financially incentivized for accuracy in an identical manner across samples (and between two treatments in the general public's survey), which mitigates the risk of social desirability bias.

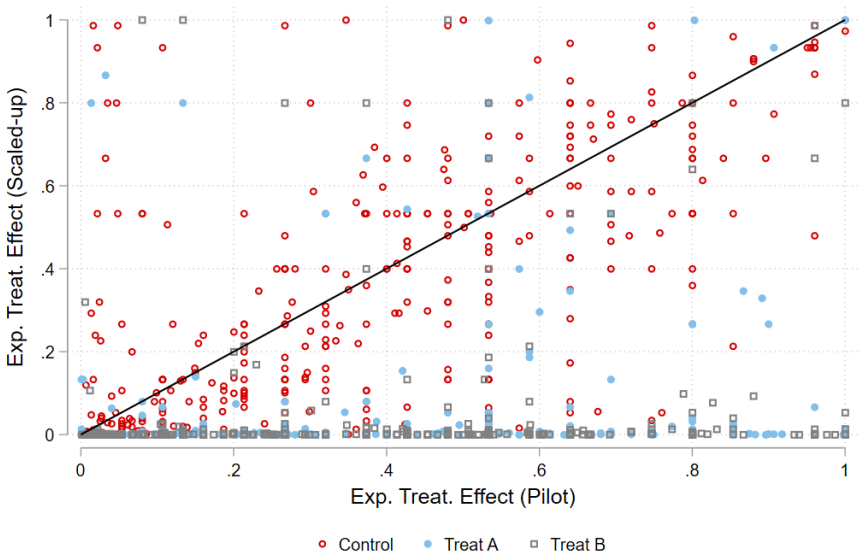
Figures 2 and 3 show that the treatment - that is, showing respondents the results of pilot trial - significantly adjusted respondents' forecast of the full-scale trial in both the policymaker and the general public sample. This is confirmed in the regression results reported in Tables A.2 and A.3. We also find that the two treatments had an almost identical effect in updating beliefs among the general public.

Figure 2: Belief updating among policymakers



Notes: The figure shows each individual's highest forecast treatment effect in the pilot trial (x-axis) against the corresponding highest forecast treatment effect in the scaled-up trial (y-axis). Sample: policymakers, N=123

Figure 3: Belief updating among general public



Notes: The figure shows each individual's highest forecast treatment effect in the pilot trial (x-axis) against the corresponding highest forecast treatment effect in the scaled-up trial (y-axis). Sample: general public, N=1,200

Expected effects at scale A next consideration is what to expect at scale. One crucial question for the experimental research agenda in both policy and academic circles relates to the scale-up problem: can this idea work at scale? In its simplest form, this question relates to the proliferation of a policy from a small group to a larger group in more diverse situations (List, 2024b). An insight that emerges from our analysis is that in the control group, 64% of policymakers and 60% of Americans expect the efficacy of financial incentives to be much less when implemented at scale.[12] This suggests that most individuals have an intuition that even when they expect an intervention may be effective on a small scale, its efficacy will diminish when scaled. In the scaling literature, this is denoted as the voltage effect (List, 2022) and our results provide novel descriptive evidence that most individuals have an intuition for its existence and import.

Obfuscated posterior beliefs. In the previous analysis, we demonstrated that both policymakers and the public internalize new information from an RCT and adjust their posterior beliefs accordingly. Next, we explore whether this belief updating extends to respondents' broader questioning of the efficacy of such interventions in other contexts. Encouraging critical thinking and questioning is an important step, as some respondents may not fully internalize the new information beyond the specific RCT context. To investigate this issue, we included an obfuscated posterior belief question in the final module of the survey, embedded among the demographic questions. This time, we used a categorical variable as a

[12]13% of treated policymakers and 5% in Treatment A and 4% in Treatment B in the general public sample expected the treatment effect to be higher at scale.

more stringent measure of belief updating.[13] Tables A.4 and A.5 show that both samples — policymakers and the general public — internalize the treatment and become more likely to believe that seeding incentives are either ineffective or uncertain in their efficacy at increasing participation rates in savings products. Interestingly, among policymakers, the strongest moderator of the treatment effect is whether their organization frequently uses seeding incentives in its programs. This could be due to policymakers' prior experiences, where they may have confidence in the results of another unpublished robust evaluation proving the efficacy of such interventions, or it could be explained by cognitive dissonance. The latter would be in line with the results on the general public sample, where belief in the importance of 529 plans and expectations of pilot trial take-up rates serve as significant moderators of the measured treatment effect.

4.3.2 Preferences for public spending on experimental evaluations and scaling

We now examine how learning from a pilot trial impacts respondents' stated preferences for allocating public resources. While previous studies have focused on whether policymakers or politicians would adopt a successful intervention in their programs or services, our analysis extends to understanding the trade-offs policymakers face in allocating resources among evaluations, scaling interventions, and maintaining business-as-usual activities.

In Table A.6, we observe that learning about the ineffectiveness of an intervention piloted via an RCT leads policymakers to both i) shift resource allocation away from scaling the intervention and ii) testing the same intervention on another sample population. This result indicates that policymakers internalized the pilot findings, highlighting the value of small-scale evaluations in preventing the expansion of ineffective interventions. In lieu of those activities, empirical results in Table A.6 exhibit a higher preference for reallocating resources towards launching new RCTs to test different interventions, rather than maintaining the status quo by allocating resources towards business as usual activities. This suggests that when confronted with trade-offs, policymakers may be inclined to substitute scaling ineffective programs with conducting more RCTs. In the general public sample (Table A.7), we see a similar pattern, though with one notable difference: respondents also increased their allocation of resources toward business-as-usual activities, in addition to launching new trials. This variation may reflect a higher baseline demand for policy experimentation among policymakers, who may prioritize new trials over maintaining existing activities.

There may be instances when policymakers face pressure to implement robust evaluations with dedicated budgets.[14] In a another question, we asked respondents to allocate

[13]The question asked: "*While the efficacy of seeding incentives may depend on several factors, including the amount of the incentive, conditionality rules, and other context-specific components, on average, how effective do you think seeding incentives are in increasing account opening rates (e.g., 529 plans, retirement plans, or others) in your state?*". Responses were collected using a seven-point Likert scale in the general public survey and a five-point Likert scale in the policymakers' survey, for design simplicity. In the analysis, responses were condensed into three categories: ineffective, effective, and unsure or neutral.

[14]Initiatives exist to encourage greater use of evaluation methodologies, with some linking federal funding

a hypothetical $100K budget between evaluating a new RCT or a legacy program, highlighting the trade-offs between the two. Interestingly, nearly all policymakers chose to split the budget equally between the two evaluations. In contrast, respondents from the general public sample, particularly those in the treatment group, became more averse to evaluating a legacy program, opting instead to allocate a significantly larger share of the budget to a new trial.

Scalability Vital Signs Common pitfalls when scaling interventions that appeared promising in pilots include "false positives" and concerns about representativeness of both the population and context (List, 2022). In the previous section, we noted that many respondents intuitively expect most interventions to be less effective at scale (the Voltage Effect). Among policymakers in the control group, 76% allocated an equal or greater share of resources to scaling compared to replicating the trial on a different sample. Similarly, 60% of the general public control group held this preference. Although our policymaker sample was too small for conclusive analysis, an examination of the general public sample revealed no single covariate that strongly correlated with this preference. This suggests that many respondents who did not receive the treatment (the expectation-correcting information) would allocate comparable or greater resources to scaling than to replicating the trial on another sample.

Another critical factor in assessing scalability is considering spillovers, or general equilibrium effects — whether an intervention that benefits some might have unintended negative spillover effects on others, or an intervention that generates positive behavioral change in one domain leads to worse outcomes in another (e.g., increasing college savings at the expense of lower retirement savings). To elicit beliefs about spillover effects, we asked all respondents to rate the likelihood of four potential spillover effects, two positive and two negative, presented in random order.

Empirical results, presented in Tables A.10 and A.11, show that most policymakers are more optimistic about positive spillovers, with 65% in the control group and 62% in the treatment group assigning a higher likelihood to positive effects. This pattern is remarkably similar among the general public sample, with 67% in the control group and 57% in the treatment group expecting positive spillovers to outweigh negative ones. We find that the difference between the control and treatment groups in the general public sample is statistically significant, suggesting that learning from pilot evaluations can induce more critical views of scalability indicators. Interestingly, among policymakers, optimism about positive spillover effects is correlated with their expectation of remaining in their current role for more than three years. In the general public sample, higher optimism is moderated by expectations of the pilot's treatment effectiveness and a stronger belief in the program's benefits. These findings suggest that cognitive dissonance may be a key moderator in learning effects from pilot evaluations. More broadly, they add new empirical content to the role that spillovers play in the scale up problem (List, 2022).

to conducting evaluations. For example, the U.S. Department of Labor's Employment and Training Administration (ETA) offers competitive grants that often require a portion of the funding to be spent on evaluation activities.

4.3.3 Demand for policy experiments

Our previous results indicate that learning about a pilot trial outcome that is at odds with expected results did not generate more experimentation aversion among policymakers. Yet, we could not exclude that there is a gap between stated and revealed preferences given the experimenter demand effects of our survey. To measure revealed preferences, and test how the treatment might have affected the demand for policy experimentation, we leveraged the upcoming conference in the policymakers' survey by asking respondents to sign up for an in-person meeting organized by conference planners. The meeting aimed to explore the feasibility of a national research consortium, where states and corporate affiliates could collaborate on data analysis and experimentation with affiliated university researchers to enhance knowledge sharing for evidence-based policymaking. We find that the treatment marginally reduces respondents' willingness to sign up for the event, with control group respondents reporting an average of 67% interest, compared to 60% in the treatment group (p=0.0912).

In the general public sample, prior belief elicitation indicates that 59% of respondents supported or strongly supported the use of RCTs to evaluate public policy efficacy. This high level of support remained largely unchanged when respondents learned about the pilot trial results (Tables A.13 and A.14), and even when they received additional information explaining the importance of RCTs, especially when results deviate from expectations (i.e., our second treatment). This finding aligns with Dur et al. (2024), who observed broad voter support for policy experiments.

As an additional indicator of interest in policy experimentation, we tracked respondents' click-through rates on two links provided at the end of the survey, directing them to MIT and University of Chicago's online introductory materials on RCTs. Among policymakers, the aversion to experimentation observed in their willingness to attend the in-person event extended to their interest in learning about conducting policy experiments. Only 4% of treated respondents, compared to 11% of the control group (p=0.0612), clicked on the links. However, we did not observe this trend in the general public sample, where approximately 1% of respondents across groups clicked on the links. We confirm these results by showing that the general public's demand for policy experiments was unaffected by either treatment as further validated by the outcomes of an incentive-compatible charitable donation question (see Online Appendix). We view this set of results as an important dissimilarity between policymakers and the general public.

4.3.4 Trust in institutions

In the previous section, we presented evidence suggesting that policymakers' demand for overall policy experimentation may decline after learning the unexpected results of a pilot trial. In contrast, the general public's demand for policy experiments remains high, and unaffected, by unanticipated outcomes. This contrast raises an important question: why is

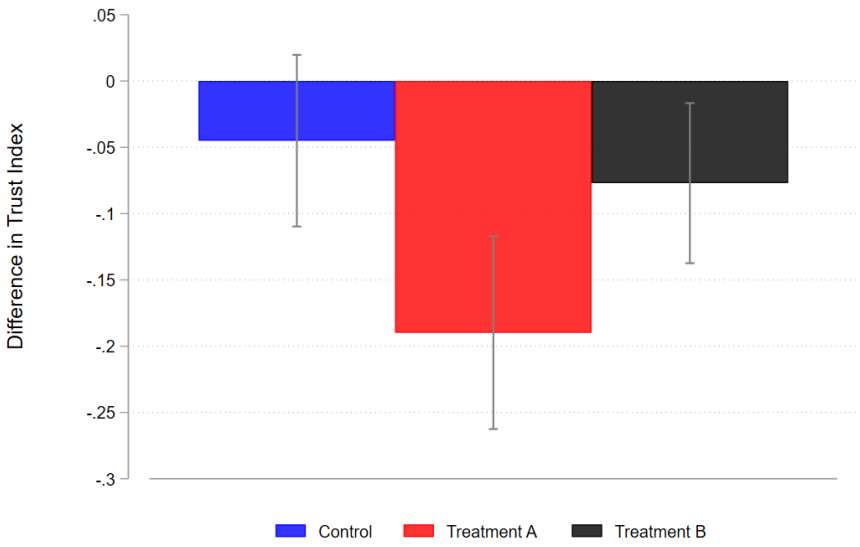
there a stated and revealed preferences gap suggesting that policymakers develop an aversion to experimentation, despite public support for such initiatives? One plausible explanation is that policymakers fear that disappointing evaluation results could negatively impact their perceived performance, potentially harming public trust in their work or citizens' willingness to engage with the program they manage.

We investigate this mechanism by examining the general public's trust in their state treasurer's office (the trial implementation agency) at the beginning (priors) and the end (posteriors) of the survey. Trust is measured along three key dimensions — ability, benevolence, and integrity — using validated survey instruments from public administration and management science literature. These dimensions capture whether the institution is seen as competent (ability), acts in the public's interest (benevolence), and prioritizes transparency and fairness (integrity). As shown in Figure 4 and Tables A.15 to A.20, learning about the pilot trial results significantly erodes public trust in the competence and integrity of the agency responsible for trial implementation. However, perceptions of the agency's benevolence are not significantly affected by the treatment.

To explore whether these negative effects can be mitigated, we included an additional treatment where respondents received a brief explanation of the importance of trial results, even when outcomes deviate from prior expectations. This explanation also highlighted the complexity of running rigorous trials and the policymakers' commitment to using evidence to create policies that actually help people. The results, displayed in Figure 4, indicate that this additional information significantly mitigated the negative impact on perceptions of the agency's competence and the tabled results even reveal increased perceptions of benevolence compared to the control group (as well as the group that only saw the trial results). Both treatments (with and without the additional explanation), however, had a similar negative effect on perceptions of the agency's integrity.

These findings suggest a potential backfire effect from transparent, evidence-based policymaking. While citizens do not oppose experimentation per se, they may interpret disappointing trial outcomes as a sign of government incompetence, rather than appreciating the government's commitment to rigorous evaluation and its capacity to handle complex assessments. Further, even in the presence of a mitigating educational intervention, perceptions of the institutional transparency, honesty and fairness (i.e., its integrity) may be negatively affected. Our data highlight the delicate balance policymakers must strike between transparency and maintaining public trust, especially when results are unexpected or undesirable.

Figure 4: Δ Trust Index: Competence, Benevolence, Integrity



Notes: The figure shows the difference in the average trust score across the three trust beliefs dimensions among respondents randomly allocated to the control group, the Treatment A that learned about the trial results, and the Treatment B that learned about the trial results and the importance of robust evaluations regardless of the results.

4.3.5 Evidence of additional mechanisms and robustness checks

Partisan motivation. Previous experimental survey studies in political science have documented the presence of a "cheerleading effect," where respondents are not necessarily resistant to new information but interpret it in a way that aligns with their political ideology (Gaines et al., 2007). In our experiment, it is plausible that this tendency moderated the treatment effects, potentially influencing how respondents processed the pilot trial results.

To explore this possibility, in the final specification of Tables A.15 to A.20, we introduce a variable indicating whether the respondent's political ideology aligns with that of their state's governor. To measure this, we asked respondents to report both their party affiliation and that of their governor, considering that some may be unaware of their governor's affiliation, which we interpret as an indicator of low importance placed on political alignment. Our findings reveal that this political alignment dummy is consistently and strongly correlated with higher trust in the state treasurer's office across all three dimensions—competence, benevolence, and integrity. However, its inclusion in the regression models does not systematically affect our treatment effect estimates. This suggests that while partisanship is associated with baseline trust in government institutions, it does not moderate the effects of learning unexpected trial results on trust.

Pilot Results Effect on Institutional Trust. Our analysis reveals significant treatment effects on the erosion of trust in the respondents' state governor, our pre-registration

measure of spillover effect. Specifically, learning about the disappointing trial pilot results led to a decrease in respondents' beliefs regarding the competence and integrity of staff employed in their state governor's office. Interestingly, we did not observe any negative effects on perceptions of benevolence, suggesting that respondents differentiated between the three dimensions of institutional trust — competence, benevolence, and integrity — rather than responding uniformly across all measures.

Regarding the mitigation effects of the information provision treatment, we find that it successfully counteracted the backfire effect on perceived competence, leading to a more favorable evaluation of the governor's staff in this area. Moreover, the information provision treatment significantly increased respondents' beliefs in the benevolence of the governor's staff compared to the primary treatment. Both treatments (the trial results alone and the trial results with information), however, led to a significant reduction in trust in the governor's office overall.

These findings suggest that respondents may generalize their trust judgments across institutions, meaning that trust erosion in one organization, such as the state treasurer's office, can spill over to other related institutions, such as the governor's office. Our results highlight the broader institutional consequences of sharing unexpected trial results, as it can affect trust in governance beyond the immediate agency responsible for the trial.

Framing effects. In the survey experiment conducted on the representative sample of Americans, we chose to anonymize the trial implementation agency and the state where the trial was conducted. This decision was made to prevent potential biases stemming from partisan opinions about state politics and to maintain consistency between the general public's survey and that of the policymakers. Respondents in the general public survey were asked to imagine the trial took place in their state after reading the trial information page.

To assess whether this anonymized framing affected our treatment effect estimates, we implemented an additional test. Specifically, we randomly assigned half of the respondents from Illinois to a version of the trial information page that explicitly stated the trial was conducted by the Illinois State Treasurer's Office. This allows us to test if identifying the agency and state influences how respondents answered the questions or reacted to the treatments. Importantly, the state of residence question was placed in the first module of the survey, enabling us to implement a conditional randomization only for those who reported living in Illinois.

Of the 45 Illinois respondents, 22 were shown the identifiable version of the trial information. In Table B.9, we reanalyze the main outcomes for this Illinois subsample, including a dummy variable indicating whether the respondent saw the identifiable version of the survey. We also include an interaction term between this dummy and the treatments. Our analysis reveals no significant differences between those who saw the anonymous version and those who saw the identifiable version of the survey, suggesting that the framing did not bias our treatment effect estimates. This robustness check supports the validity of our original design, ensuring that anonymizing the trial did not meaningfully affect respondents'

reactions to the treatment. Of course, our test is over very small samples and with a larger sample size this treatment might have import.

Multiple hypothesis testing. Lastly, we applied the multiple hypothesis testing (MHT) correction following the methodology proposed by List et al. (2019) to account for the potential inflation of Type I errors when evaluating multiple outcomes. Specifically, we tested the three main dimensions of trust in the government agency responsible for implementing the trial: competence, benevolence, and integrity.[15]

We conducted a regression analysis on the differences between respondents' posteriors and priors for these three dimensions, across the two treatments, controlling for all covariates included in the full specification from previous estimations. The results show that the loss of trust in the implementation agency's competence remains robust and statistically significant across all MHT corrections. The estimated effect of the main treatment on the loss of trust in the agency's integrity is significant at the 10% level in all specifications, except under the most stringent Bonferroni correction. These findings reinforce our initial results, indicating that the trust erosion in the agency's competence is particularly robust, while the negative effect on integrity is weaker but still present under less conservative corrections. All other results align with the full model specifications.

5 Conclusions

This study explores the demand and supply for policy experimentation. We begin with a large-scale natural field experiment that evaluated the efficacy of time-bound small financial incentives to increase uptake of college savings accounts, a common policy implemented across several U.S. states but not previously evaluated experimentally. We find no evidence of effectiveness. We then turn to the core of the study: a survey experiment with both U.S. state policymakers and the general American public. We report several interesting insights.

First, there is considerable policymaker adaptability. In this manner, it is a positive result that policymakers adjust their beliefs and resource allocation based on new evidence. This adaptability is crucial for science-driven policy. Second, policymakers may tend to generalize from one set of results to a broader skepticism of scientific methods. This result is concerning, highlighting the potential benefits of enhancing communication about the iterative nature of science. Third, there is considerable public and policymaker alignment. The similarities in how the public and policymakers view certain results suggest a shared understanding, which can be leveraged for more effective policies. Yet, a fourth result is that at the same time there are key differences: while we observe a deterioration of support for science among policymakers, general citizens remain optimistic about scientific methods. Yet, there is a public cost for negative trial results. Our final set of results reveals that trust issues arise: policymaker trust is eroded after disappointing results are reported, indicating the

[15]The MHT results remain consistent when extending the analysis to include the three trust dimensions for the state Governor's office as well.

need for transparency and effective communication. This is a key spot for an educational role, as educating the public about the value and uncertainty of policy experiments can help to restore trust, emphasizing that learning from failures is a strength, not a weakness.

In sum, our findings suggest that while public support for policy experimentation is strong and resilient in the face of unexpected results, there is a critical need to manage both policymakers' and citizens' expectations. Policymakers' aversion to experimentation after disappointing results is a challenge to evidence-based governance, indicating the value of communication strategies that focus on the lessons trials can offer, regardless of their outcomes. Similarly, the public's declining trust in institutions after learning about unexpected results suggests that greater efforts are needed to educate citizens about the role and complexity of government-led experiments. In contrast to previous studies, which primarily focus on how policymakers adjust their beliefs after learning about positive trial results, we show that disappointing outcomes can lead to reduced enthusiasm for experimentation. This highlights the importance of preparing policymakers for all possible outcomes in an experimental setting to ensure continued commitment to evidence-based policymaking. Moreover, our study describes the potential consequences of educating the public about policy experimentation, particularly in terms of its role in improving transparency, accountability, and ultimately trust in government.

Future studies should explore further how the public's demand for policy evaluation may fluctuate when trial results do not align with their expectations or political ideologies. Additionally, research could focus on the effectiveness of educational interventions that leverage policy experiments as tools for enhancing government accountability and fostering trust among citizens. By doing so, we can better understand how to maintain public and policymaker support for experimentation, even in the face of unexpected or disappointing outcomes.

References

Agostinelli, F., C. Avitabile, and M. Bobba (2023). Enhancing human capital in children: A case study on scaling. Technical report, National Bureau of Economic Research.

Al-Ubaydli, O., J. A. List, and D. L. Suskind (2017). What can we learn from experiments? understanding the threats to the scalability of experimental results. *American Economic Review 107*(5), 282–286.

Carattini, S., R. Dur, and J. List (2024). Policy evaluation and the causal analysis of public support. *Science 386*(6721), 490–492.

DellaVigna, S., W. Kim, and E. Linos (2024). Bottlenecks for evidence adoption. *Journal of Political Economy 132*(8), 2748–2789.

DellaVigna, S. and D. Pope (2018). Predicting experimental results: who knows what? *Journal of Political Economy 126*(6), 2410–2456.

Dur, R., A. Non, P. Prottung, and B. Ricci (2024). Who's afraid of policy experiments? *The Economic Journal*, ueae090.

Gaines, B. J., J. H. Kuklinski, P. J. Quirk, B. Peyton, and J. Verkuilen (2007). Same facts, different interpretations: Partisan motivation and opinion on iraq. *The Journal of Politics 69*(4), 957–974.

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.

Herd, P. and D. P. Moynihan (2019). *Administrative burden: Policymaking by other means*. Russell Sage Foundation.

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.

Jilke, S., F. Keppeler, J. Ternovski, D. Vogel, and E. Yoeli (2024). Policy makers believe money motivates more than it does. *Scientific Reports 14*(1), 1901.

Landry, C. E., A. Lange, J. A. List, M. K. Price, and N. G. Rupp (2010). Is a donor in hand better than two in the bush? evidence from a natural field experiment. *American Economic Review 100*(3), 958–983.

Larroucau, T., I. Rios, A. Fabre, and C. Neilson (2024). College application mistakes and the design of information policies at scale.

Levitt, S. D. and J. A. List (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review 53*(1), 1–18.

List, J. A. (2022). *The voltage effect: How to make good ideas great and great ideas scale*. Crown Currency.

List, J. A. (2024a). *Experimental Economics: Theory and Practice*. University of Chicago Press.

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

List, J. A., A. M. Shaikh, and Y. Xu (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics 22*, 773–793.

Mason, L. R., Y. Nam, M. Clancy, Y. Kim, and V. Loke (2010). Child development accounts and saving for children's future: Do financial incentives matter? *Children and Youth Services Review 32*(11), 1570–1576.

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

Mazar, N., C. T. Elbaek, and P. Mitkidis (2023). Experiment aversion does not appear to generalize. *Proceedings of the National Academy of Sciences 120*(16), e2217551120.

Mobarak, A. M. (2022). Assessing social aid: the scale-up process needs evidence, too. *Nature 609*(7929), 892–894.

Sahni, N. S., S. C. Wheeler, and P. Chintagunta (2018). Personalization in email marketing: The role of noninformative advertising content. *Marketing Science 37*(2), 236–258.

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

Tummala, S., W. Zhong, and L. E. Marcil (2022). Embedding 529 college savings accounts in pediatric care: a pilot innovation. *Academic pediatrics 22*(3), 501–502.

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

Wang, S. and D. Y. Yang (2021). Policy experimentation in china: The political economy of policy learning. Technical report, National Bureau of Economic Research.

Wiswall, M. and B. Zafar (2015). Determinants of college major choice: Identification using an information experiment. *The Review of Economic Studies 82*(2), 791–824.

6 Appendix

Table A.1: Natural field experiment results

		Incentive Amount			
		\$0	\$10	\$50	\$100
Pilot, N=7,500	Claimed seed	-	0	2	1
	Opened Accounts	1	0	2	1
Full-scale, N=150,000	Claimed seed	-	11	27	23
	Opened Accounts	0	1	11	9

Notes: The table shows the number of seed claimants and account openers across randomized groups between the soft launch and the full-scale trial.

Table A.2: Belief updating among policymakers: forecast treatment effect in scaled-up trial

Exp. highest take-up rate (Scaled-up)	(1)	(2)	(3)
Treatment	-0.184***	-0.172***	-0.179***
	(0.0298)	(0.0235)	(0.0237)
Exp. highest take-up rate (Pilot)		0.456***	0.461***
		(0.0523)	(0.0520)
Constant	0.195***	0.0777***	0.152**
	(0.0211)	(0.0214)	(0.0607)
Controls	No	No	Yes
N	124	124	124
R^2	0.237	0.531	0.603

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables in model (3) include: Confident in pilot priors, Postgrad Educ, Female, Org. has used RCTs and seeding incentives for takeup, works on 529 plans, in current job 3+ years, expects to stay 3+ years, Age 35-54 or 55+.

Table A.3: Belief updating among the general public: forecast treatment effect in scaled-up trial

Exp. highest take-up rate (Scaled-up)	(1)	(2)	(3)	(4)
Treatment A	-0.299***	-0.301***	-0.303***	-0.303***
	(0.0153)	(0.0142)	(0.0142)	(0.0142)
Treatment B	-0.306***	-0.308***	-0.309***	-0.309***
	(0.0154)	(0.0142)	(0.0143)	(0.0143)
Confident in pilot priors		0.0110	0.00634	0.00588
		(0.0139)	(0.0142)	(0.0143)
Exp. highest take-up rate (Pilot)		0.303***	0.303***	0.302***
		(0.0211)	(0.0213)	(0.0213)
Constant	0.345***	0.228***	0.277***	0.276***
	(0.0109)	(0.0130)	(0.0337)	(0.0339)
Controls	No	No	Yes	Yes
N	1,200	1,200	1,200	1,200
R^2	0.302	0.406	0.413	0.413

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables in models (3) and (4) include: Heard of the program before, College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH income, believes the program helps. Model (4) also includes number of seconds spent reading the trial info page.

Table A.4: Belief updating among policymakers: obfuscated posterior

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Ineffective	Unsure	Ineffective	Unsure	Ineffective	Unsure	Ineffective	Unsure
Treatment	0.549	0.488	0.830*	0.0934	0.822*	0.338	0.801	-0.0220
	(0.439)	(0.585)	(0.498)	(0.957)	(0.497)	(1.011)	(0.501)	(0.989)
Confident in pilot priors			-0.158	1.501	-0.152	1.546	-0.182	1.514
			(0.483)	(0.995)	(0.483)	(0.991)	(0.484)	(0.996)
Resp's org. used RCTs before			-0.137	-2.801*	-0.136	-3.014**	-0.177	-3.040*
			(0.526)	(1.450)	(0.526)	(1.485)	(0.532)	(1.582)
Resp's org. uses seeding incentives for takeup			-1.184**	-4.095***	-1.185**	-4.495***	-1.152*	-4.192***
			(0.588)	(1.383)	(0.591)	(1.563)	(0.591)	(1.439)
Resp. works on 529 plans			0.475	-1.857*	0.483	-1.848*	0.506	-1.868*
			(0.600)	(1.056)	(0.600)	(1.082)	(0.601)	(1.075)
Has had current job for at least 3 years			0.977	1.531	0.972	1.256	0.936	1.667
			(0.664)	(1.063)	(0.664)	(1.105)	(0.669)	(1.139)
Expects to keep current job for 3+ years			-1.306*	-0.928	-1.315*	-1.199	-1.277*	-1.193
			(0.685)	(1.391)	(0.686)	(1.392)	(0.685)	(1.467)
Exp. highest take-up rate (Pilot)					-0.0365	2.212		
					(1.077)	(2.465)		
Constant	-1.299***	-1.992***	-0.129	2.789	-0.127	2.749	0.0155	3.503
	(0.326)	(0.435)	(1.342)	(2.194)	(1.360)	(2.260)	(1.384)	(2.627)
Controls	No	No	Yes	Yes	Yes	Yes	Yes	Yes
N	123	123	123	123	123	123	123	123

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: Postgrad Educ, Female, Age 35-54 or 55+. Models (7) and (8) also include number of seconds spent reading the trial info page.

Table A.5: Belief updating among the general public: obfuscated posterior

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Ineffective	Unsure	Ineffective	Unsure	Ineffective	Unsure	Ineffective	Unsure
Treatment A	1.944***	0.895***	1.965***	0.924***	2.064***	1.020***	2.064***	1.023***
	(0.179)	(0.233)	(0.180)	(0.234)	(0.187)	(0.240)	(0.187)	(0.241)
Treatment B	1.870***	0.995***	1.876***	1.008***	1.931***	1.106***	1.931***	1.125***
	(0.180)	(0.228)	(0.181)	(0.230)	(0.188)	(0.236)	(0.188)	(0.237)
Confident in pilot priors			-0.331**	-0.327	-0.280	-0.253	-0.277	-0.211
			(0.163)	(0.234)	(0.171)	(0.242)	(0.172)	(0.243)
Exp. highest take-up rate (Pilot)			-0.553**	-0.898**	-0.387	-0.897**	-0.378	-0.864**
			(0.246)	(0.351)	(0.255)	(0.359)	(0.256)	(0.359)
Heard of the program before					0.201	0.0987	0.199	0.0835
					(0.151)	(0.205)	(0.151)	(0.206)
Believes the program helps					-0.443***	-0.357***	-0.442***	-0.347***
					(0.0669)	(0.0890)	(0.0670)	(0.0892)
Constant	-1.647***	-1.994***	-1.370***	-1.602***	1.101***	0.909*	1.109***	1.080**
	(0.143)	(0.166)	(0.170)	(0.207)	(0.410)	(0.539)	(0.413)	(0.546)
Controls	No	No	No	No	Yes	Yes	Yes	Yes
N	1,200	1,200	1,200	1,200	1,200	1,200	1,200	1,200

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: Heard of the program before, College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH income. Models (7) and (8) also include number of seconds spent reading the trial info page.

Table A.6: Policymakers' resource allocation preferences

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	BAU	BAU	Same RCT, diff sample	Same RCT, diff sample	New RCT	New RCT	Scale-up	Scale-up
Treatment	0.0353	0.0465	-0.0763**	-0.0874**	0.104***	0.120***	-0.0635**	-0.0788**
	(0.0444)	(0.0471)	(0.0342)	(0.0361)	(0.0362)	(0.0377)	(0.0304)	(0.0324)
Confident in pilot priors		0.0285		0.00704		-0.0335		-0.00207
		(0.0463)		(0.0355)		(0.0371)		(0.0319)
Resp's org. used RCTs before		-0.0638		0.00384		0.0123		0.0477
		(0.0513)		(0.0394)		(0.0412)		(0.0354)
Resp's org. uses seeding incentives for takeup		0.00807		-0.0467		0.0233		0.0154
		(0.0559)		(0.0429)		(0.0448)		(0.0385)
Resp. works on 529 plans		0.0823		0.0136		-0.0678		-0.0281
		(0.0557)		(0.0428)		(0.0447)		(0.0384)
Has had current job for at least 3 years		-0.0376		0.0337		0.0322		-0.0283
		(0.0554)		(0.0425)		(0.0444)		(0.0381)
Expects to keep current job for 3+ years		-0.0548		0.0719		-0.0225		0.00552
		(0.0729)		(0.0559)		(0.0585)		(0.0502)
Constant	0.298***	0.394***	0.304***	0.263***	0.205***	0.0930	0.194***	0.250***
	(0.0315)	(0.113)	(0.0243)	(0.0867)	(0.0257)	(0.0906)	(0.0216)	(0.0778)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	131	131	131	131	131	131	131	131
R^2	0.005	0.075	0.037	0.114	0.060	0.158	0.033	0.093

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: Postgrad Educ, Female, Age 35-54 or 55+ and number of seconds spent reading the trial info page.

Table A.7: General public's resource allocation preferences

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	BAU	BAU	Same RCT, diff sample	Same RCT, diff sample	New RCT	New RCT	Scale-up	Scale-up
Treatment A	0.0614***	0.0593***	-0.0577***	-0.0545***	0.0895***	0.0883***	-0.0932***	-0.0931***
	(0.0176)	(0.0174)	(0.0135)	(0.0132)	(0.0173)	(0.0170)	(0.0139)	(0.0138)
Treatment B	0.0696***	0.0656***	-0.0737***	-0.0682***	0.115***	0.110***	-0.111***	-0.108***
	(0.0176)	(0.0175)	(0.0135)	(0.0133)	(0.0173)	(0.0172)	(0.0139)	(0.0139)
Confident in pilot priors		-0.0381**		-6.29e-05		0.0102		0.0279**
		(0.0175)		(0.0133)		(0.0172)		(0.0139)
Exp. highest take-up rate (Pilot)		-0.0513*		0.0700***		-0.0740***		0.0553***
		(0.0262)		(0.0199)		(0.0257)		(0.0208)
Heard of the program before		0.00245		0.00483		0.0119		-0.0192
		(0.0154)		(0.0117)		(0.0151)		(0.0123)
Constant	0.228***	0.306***	0.266***	0.0906***	0.233***	0.430***	0.273***	0.173***
	(0.0124)	(0.0417)	(0.00955)	(0.0316)	(0.0122)+	(0.0408)	(0.00985)	(0.0331)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
N	1,200	1,200	1,200	1,200	1,200	1,200	1,200	1,200
R^2	0.015	0.051	0.027	0.082	0.039	0.079	0.058	0.082

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH income and seconds spent reading the trial info page.

Table A.8: Policymakers' resource allocation preferences for evaluations

Budget share allocated to evaluating a legacy program (instead of launching a new trial)	(1)	(2)
Treatment	0.0571	0.0370
	(0.0377)	(0.0397)
Confident in pilot priors		0.0652*
		(0.0389)
Resp's org. used RCTs before		-0.0195
		(0.0435)
Resp's org. uses seeding incentives for takeup		-0.00695
		(0.0473)
Resp. works on 529 plans		0.0252
		(0.0472)
Has had current job for at least 3 years		-0.0410
		(0.0468)
Expects to keep current job for 3+ years		-0.0383
		(0.0617)
Constant	0.469***	0.515***
	(0.0270)	(0.0947)
Controls	No	Yes
N	133	133
R^2	0.017	0.105

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: Postgrad Educ, Female, Age 35-54 or 55+ and number of seconds spent reading the trial info page.

Table A.9: General public's resource allocation preferences for evaluations

Budget share allocated to evaluating a legacy program (instead of launching a new trial)	(1)	(2)	(3)	(4)
Treatment A	-0.0416***	-0.0418***	-0.0403***	-0.0402***
	(0.0143)	(0.0143)	(0.0143)	(0.0142)
Treatment B	-0.0307**	-0.0306**	-0.0274*	-0.0271*
	(0.0144)	(0.0144)	(0.0144)	(0.0144)
Confident in pilot priors		0.00658	0.00222	0.00442
		(0.0141)	(0.0143)	(0.0144)
Exp. highest take-up rate (Pilot)		0.0269	0.0168	0.0189
		(0.0214)	(0.0215)	(0.0215)
Heard of the program before			-0.0132	-0.0140
			(0.0126)	(0.0126)
Believes the program helps			0.0121**	0.0126**
			(0.00543)	(0.00544)
Constant	0.492***	0.481***	0.450***	0.457***
	(0.0101)	(0.0132)	(0.0339)	(0.0341)
Controls	No	No	Yes	Yes
N	1,200	1,200	1,200	1,200
R^2	0.007	0.009	0.030	0.033

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH incomes. Model (4) also includes number of seconds spent reading the trial info page.

Table A.10: Policymaker's belief of spillovers at scale

Believes negative spillovers are more likely than positive when scaled	(1)	(2)	(3)
Treatment	0.0331	0.0337	0.0334
	(0.0810)	(0.0836)	(0.0836)
Confident in pilot priors		-0.102	-0.113
		(0.0827)	(0.0835)
Resp's org. used RCTs before		0.100	0.0939
		(0.0957)	(0.0959)
Resp's org. uses seeding incentives for takeup		0.0212	0.0264
		(0.104)	(0.105)
Resp. works on 529 plans		-0.0225	-0.0186
		(0.104)	(0.104)
Has had current job for at least 3 years		0.00714	-0.00431
		(0.102)	(0.103)
Expects to keep current job for 3+ years		-0.257*	-0.256*
		(0.137)	(0.137)
Constant	0.347***	0.654***	0.679***
	(0.0571)	(0.200)	(0.202)
Controls	No	Yes	Yes
N	143	143	143
R^2	0.001	0.080	0.087

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: Postgrad Educ, Female, Age 35-54 or 55+. Model (3) also include number of seconds spent reading the trial info page.

Table A.11: General public's belief of spillovers at scale

Believes negative spillovers are more likely than positive when scaled	(1)	(2)	(3)	(4)
Treatment A	0.0854**	0.0870**	0.0859**	0.0855**
	(0.0344)	(0.0341)	(0.0334)	(0.0334)
Treatment B	0.110***	0.107***	0.100***	0.0997***
	(0.0345)	(0.0342)	(0.0337)	(0.0336)
Confident in pilot priors		-0.0890***	-0.0647*	-0.0702**
		(0.0336)	(0.0336)	(0.0337)
Exp. highest take-up rate (Pilot)		-0.219***	-0.180***	-0.186***
		(0.0508)	(0.0503)	(0.0503)
Heard of the program before			0.136***	0.138***
			(0.0297)	(0.0296)
Believes the program helps			-0.0703***	-0.0715***
			(0.0127)	(0.0127)
Constant	0.333***	0.436***	0.800***	0.781***
	(0.0244)	(0.0314)	(0.0795)	(0.0800)
Controls	No	No	Yes	Yes
N	1,200	1,200	1,200	1,200
R^2	0.009	0.032	0.079	0.082

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH incomes. Model (4) also includes number of seconds spent reading the trial info page.

Table A.12: Policymakers' revealed preferences for policy experiments

Signed up to conduct more RCTs	(1)	(2)	(3)
Treatment	-0.0716	-0.0847	-0.0927*
	(0.0534)	(0.0540)	(0.0545)
Confident in pilot priors		0.0532	0.0460
		(0.0532)	(0.0536)
Resp's org. used RCTs before		-0.0202	-0.0268
		(0.0578)	(0.0582)
Resp's org. uses seeding incentives for takeup		0.00881	0.0133
		(0.0626)	(0.0628)
Resp. works on 529 plans		0.0787	0.0833
		(0.0630)	(0.0632)
Has had current job for at least 3 years		-0.0410	-0.0530
		(0.0622)	(0.0633)
Expects to keep current job for 3+ years		0.202**	0.202**
		(0.0787)	(0.0787)
Constant	0.675***	0.485***	0.514***
	(0.0374)	(0.139)	(0.142)
Controls	No	Yes	Yes
N	114	114	114
R^2	0.016	0.170	0.178

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH incomes. Model (3) also includes number of seconds spent reading the trial info page.

Table A.13: General public's support for policy experiments

	(1)	(2)	(3)	(4)
Support for RCTs in public policy	Posteriors	Posteriors	Posteriors	Posteriors
Treatment A	-0.0791	-0.0786	-0.0780	-0.0783
	(0.0632)	(0.0632)	(0.0627)	(0.0627)
Treatment B	-0.0183	-0.0156	-0.00106	-0.00176
	(0.0676)	(0.0676)	(0.0670)	(0.0670)
Priors	0.645***	0.643***	0.624***	0.623***
	(0.0233)	(0.0234)	(0.0236)	(0.0236)
Confident in pilot priors		0.0554	0.000625	-0.00344
		(0.0674)	(0.0681)	(0.0682)
Exp. highest take-up rate (Pilot)		-0.0605	-0.117	-0.121
		(0.106)	(0.105)	(0.105)
Heards of the program before			-0.0400	-0.0384
			(0.0558)	(0.0558)
Believes the program helps			0.141***	0.140***
			(0.0284)	(0.0284)
Constant	1.851***	1.870***	1.324***	1.313***
	(0.123)	(0.130)	(0.180)	(0.180)
Controls	No	No	Yes	Yes
N	1,200	1,200	1,200	1,200
R^2	0.499	0.500	0.516	0.516
Test: Treat A = Treat B	0.3891	0.3725	0.2727	0.2752

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH income. Model (4) also includes number of seconds spent reading the trial info page.

Table A.14: General public's support for policy experiments

	(1)	(2)	(3)	(4)
Support for RCTs in public policy	Δ beliefs	Δ beliefs	Δ beliefs	Δ beliefs
Treatment A	-0.0808	-0.0810	-0.0726	-0.0724
	(0.0720)	(0.0721)	(0.0721)	(0.0722)
Treatment B	0.0289	0.0247	0.0453	0.0457
	(0.0760)	(0.0758)	(0.0765)	(0.0765)
Confident in pilot priors		-0.0875	-0.102	-0.0998
		(0.0758)	(0.0784)	(0.0783)
Exp. highest take-up rate (Pilot)		0.0140	-0.0314	-0.0291
		(0.119)	(0.120)	(0.120)
Heard of the program before			-0.0573	-0.0582
			(0.0652)	(0.0652)
Believes the program helps			0.0579**	0.0584**
			(0.0288)	(0.0288)
Constant	0.248***	0.263***	0.0564	0.0646
	(0.0462)	(0.0652)	(0.175)	(0.177)
Controls	No	No	Yes	Yes
N	1,200	1,200	1,200	1,200
R^2	0.002	0.003	0.014	0.014

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH income. Model (4) also includes number of seconds spent reading the trial info page.

Table A.15: Trust: Competence

	(1)	(2)	(3)	(4)
Competence of own state Treasurer's Office	Posteriors	Posteriors	Posteriors	Posteriors
Treatment A	-0.207***	-0.206***	-0.205***	-0.193***
	(0.0691)	(0.0692)	(0.0680)	(0.0678)
Treatment B	-0.0489	-0.0421	-0.0327	-0.0321
	(0.0656)	(0.0656)	(0.0638)	(0.0633)
Priors	0.802***	0.799***	0.760***	0.753***
	(0.0188)	(0.0190)	(0.0206)	(0.0210)
Confident in pilot priors		0.130*	0.0712	0.0599
		(0.0741)	(0.0720)	(0.0731)
Exp. highest take-up rate (Pilot)		-0.105	-0.145	-0.128
		(0.103)	(0.102)	(0.102)
Heard of the program before			0.0923	0.0976*
			(0.0580)	(0.0585)
Believes the program helps			0.198***	0.187***
			(0.0286)	(0.0286)
Republican				-0.148*
				(0.0895)
Democrat				-0.171**
				(0.0849)
State Gov is same political party				0.332***
				(0.0791)
Constant	0.744***	0.765***	-0.155	-0.119
	(0.0949)	(0.105)	(0.159)	(0.161)
Controls	No	No	Yes	Yes
N	1,200	1,200	1,200	1,200
R^2	0.611	0.612	0.633	0.639
Test: Treat A = Treat B	0.0291	0.0232	0.0154	0.0225

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH income and number of seconds spent reading the trial info page.

Table A.16: Trust: Competence

	(1)	(2)	(3)	(4)
Competence of own state Treasurer's Office	Δ beliefs	Δ beliefs	Δ beliefs	Δ beliefs
Treatment A	-0.221***	-0.220***	-0.222***	-0.211***
	(0.0724)	(0.0725)	(0.0720)	(0.0718)
Treatment B	-0.0409	-0.0370	-0.0366	-0.0354
	(0.0683)	(0.0684)	(0.0681)	(0.0678)
Confident in pilot priors		0.0667	0.0147	0.00132
		(0.0773)	(0.0780)	(0.0794)
Exp. highest take-up rate (Pilot)		-0.104	-0.119	-0.102
		(0.107)	(0.107)	(0.108)
Heard of the program before			0.0841	0.0926
			(0.0617)	(0.0625)
Believes the program helps			0.103***	0.0927***
			(0.0277)	(0.0279)
Republican				-0.0846
				(0.0952)
Democrat				-0.179*
				(0.0921)
State Gov is same political party				0.248***
				(0.0837)
Constant	-0.142***	-0.119**	-0.641***	-0.618***
	(0.0455)	(0.0604)	(0.173)	(0.175)
Controls	No	No	Yes	Yes
N	1,200	1,200	1,200	1,200
R^2	0.009	0.010	0.033	0.043
Test: Treat A = Treat B	0.0185	0.0161	0.0144	0.0199

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH income and number of seconds spent reading the trial info page.

Table A.17: Trust: Benevolence

	(1)	(2)	(3)	(4)
Benevolence of own state Treasurer's Office	Posteriors	Posteriors	Posteriors	Posteriors
Treatment A	-0.0914	-0.0918	-0.0889	-0.0868
	(0.0723)	(0.0723)	(0.0713)	(0.0709)
Treatment B	0.0964	0.0991	0.117*	0.115*
	(0.0659)	(0.0663)	(0.0656)	(0.0654)
Priors	0.792***	0.790***	0.762***	0.753***
	(0.0186)	(0.0190)	(0.0200)	(0.0201)
Confident in pilot priors		0.0754	0.0327	0.0365
		(0.0781)	(0.0779)	(0.0775)
Exp. highest take-up rate (Pilot)		0.0677	0.0214	0.0231
		(0.110)	(0.108)	(0.108)
Heard of the program before			-0.0221	-0.0277
			(0.0619)	(0.0613)
Believes the program helps			0.173***	0.168***
			(0.0269)	(0.0270)
Republican				-0.163*
				(0.0937)
Democrat				0.00691
				(0.0853)
State Gov is same political party				0.225***
				(0.0748)
Constant	0.871***	0.835***	0.159	0.160
	(0.0883)	(0.0974)	(0.165)	(0.166)
Controls	No	No	Yes	Yes
N	1,200	1,200	1,200	1,200
R^2	0.608	0.608	0.623	0.628
Test: Treat A = Treat B	0.0109	0.0097	0.0045	0.0053

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH income and number of seconds spent reading the trial info page.

Table A.18: Trust: Benevolence

	(1)	(2)	(3)	(4)
Benevolence of own state Treasurer's Office	Δ beliefs	Δ beliefs	Δ beliefs	Δ beliefs
Treatment A	-0.0997	-0.100	-0.0944	-0.0962
	(0.0767)	(0.0767)	(0.0760)	(0.0758)
Treatment B	0.118*	0.117*	0.135*	0.134*
	(0.0685)	(0.0689)	(0.0690)	(0.0689)
Confident in pilot priors		-0.00829	-0.0327	-0.0283
		(0.0825)	(0.0842)	(0.0838)
Exp. highest take-up rate (Pilot)		0.0942	0.0698	0.0674
		(0.117)	(0.116)	(0.116)
Heard of the program before			-0.0192	-0.0260
			(0.0660)	(0.0655)
Believes the program helps			0.0760***	0.0724***
			(0.0269)	(0.0270)
Republican				-0.0887
				(0.101)
Democrat				0.0623
				(0.0926)
State Gov is same political party				0.0890
				(0.0798)
Constant	0.0375	0.00415	-0.176	-0.192
	(0.0474)	(0.0642)	(0.173)	(0.174)
Controls	No	No	Yes	Yes
N	1,200	1,200	1,200	1,200
R^2	0.007	0.008	0.029	0.034
Test: Treat A = Treat B	0.0052	0.0053	0.0030	0.0030

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH income and number of seconds spent reading the trial info page.

Table A.19: Trust: Integrity

	(1)	(2)	(3)	(4)
Integrity of own state Treasurer's Office	Posteriors	Posteriors	Posteriors	Posteriors
Treatment A	-0.130**	-0.129**	-0.133**	-0.124*
	(0.0651)	(0.0648)	(0.0639)	(0.0638)
Treatment B	-0.119*	-0.110*	-0.103	-0.0998
	(0.0667)	(0.0664)	(0.0650)	(0.0648)
trust_sto_integrity_prior	0.790***	0.786***	0.749***	0.743***
	(0.0166)	(0.0168)	(0.0179)	(0.0181)
Confident in pilot priors		0.194***	0.137**	0.134*
		(0.0699)	(0.0699)	(0.0696)
Exp. highest take-up rate (Pilot)		-0.0663	-0.105	-0.0988
		(0.101)	(0.0990)	(0.0983)
Heard of the program before			-0.0585	-0.0475
			(0.0556)	(0.0560)
Believes the program helps			0.176***	0.169***
			(0.0266)	(0.0266)
Republican				-0.0683
				(0.0827)
Democrat				-0.0479
				(0.0769)
State Gov is same political party				0.209***
				(0.0692)
Constant	0.802***	0.796***	-0.0438	-0.0661
	(0.0851)	(0.0944)	(0.145)	(0.147)
Controls	No	No	Yes	Yes
N	1,200	1,200	1,200	1,200
R^2	0.651	0.653	0.671	0.674
Test: Treat A = Treat B	0.8759	0.7715	0.6414	0.7050

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH income and number of seconds spent reading the trial info page.

Table A.20: Trust: Integrity

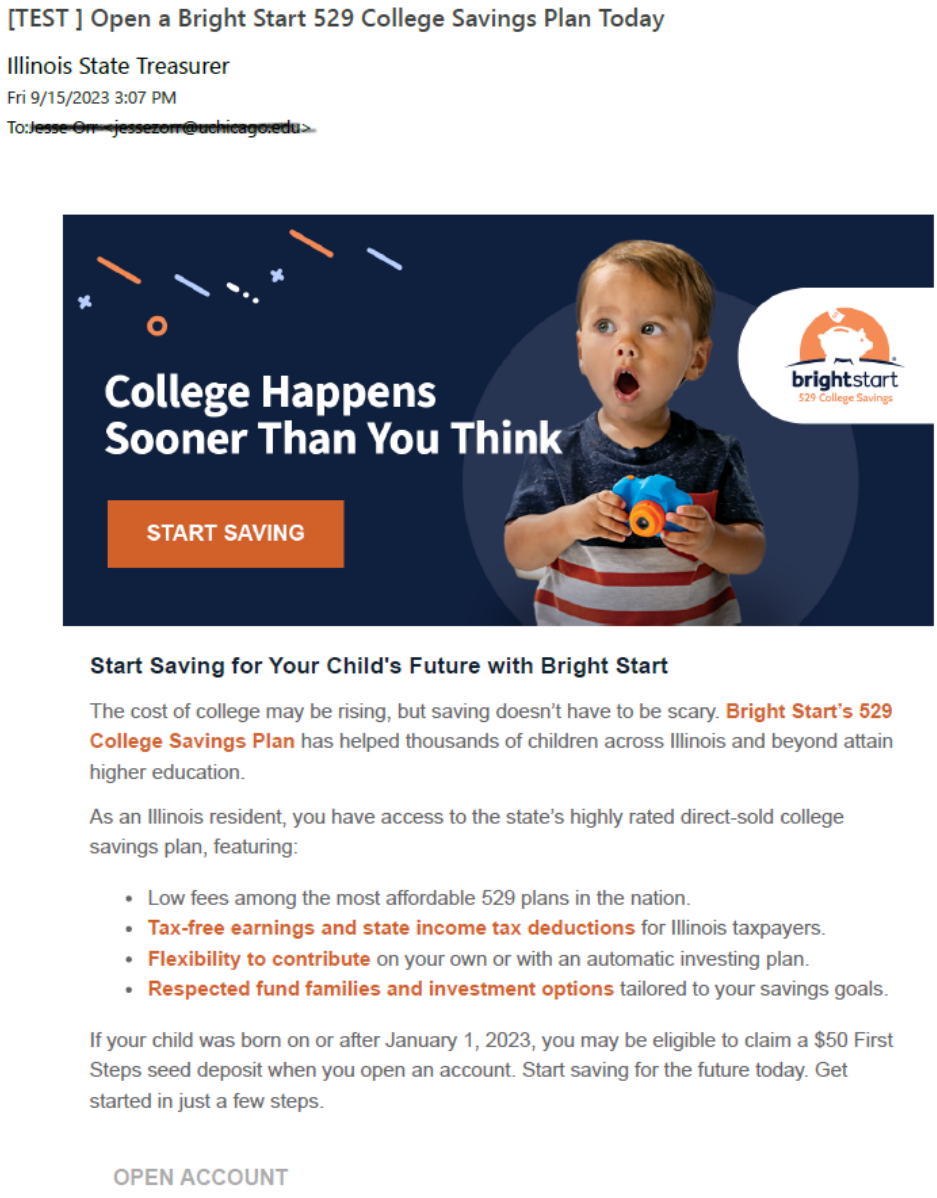
	(1)	(2)	(3)	(4)
Integrity of own state Treasurer's Office	Δ beliefs	Δ beliefs	Δ beliefs	Δ beliefs
Treatment A	-0.149**	-0.149**	-0.154**	-0.147**
	(0.0691)	(0.0690)	(0.0690)	(0.0690)
Treatment B	-0.139*	-0.132*	-0.135*	-0.132*
	(0.0711)	(0.0708)	(0.0711)	(0.0713)
Confident in pilot priors		0.135*	0.0970	0.0962
		(0.0739)	(0.0761)	(0.0763)
Exp. highest take-up rate (Pilot)		-0.0228	-0.0357	-0.0318
		(0.107)	(0.107)	(0.107)
Heard of the program before			-0.0817	-0.0661
			(0.0603)	(0.0607)
Believes the program helps			0.0732***	0.0697**
			(0.0275)	(0.0277)
Republican				0.000190
				(0.0895)
Democrat				-0.0392
				(0.0859)
State Gov is same political party				0.0899
				(0.0753)
Constant	-0.0625	-0.0863	-0.472***	-0.512***
	(0.0499)	(0.0654)	(0.160)	(0.162)
Controls	No	No	Yes	Yes
N	1,200	1,200	1,200	1,200
R^2	0.005	0.008	0.022	0.026
Test: Treat A = Treat B	0.8825	0.8141	0.7896	0.8310

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH income and number of seconds spent reading the trial info page.

7 Online Appendix

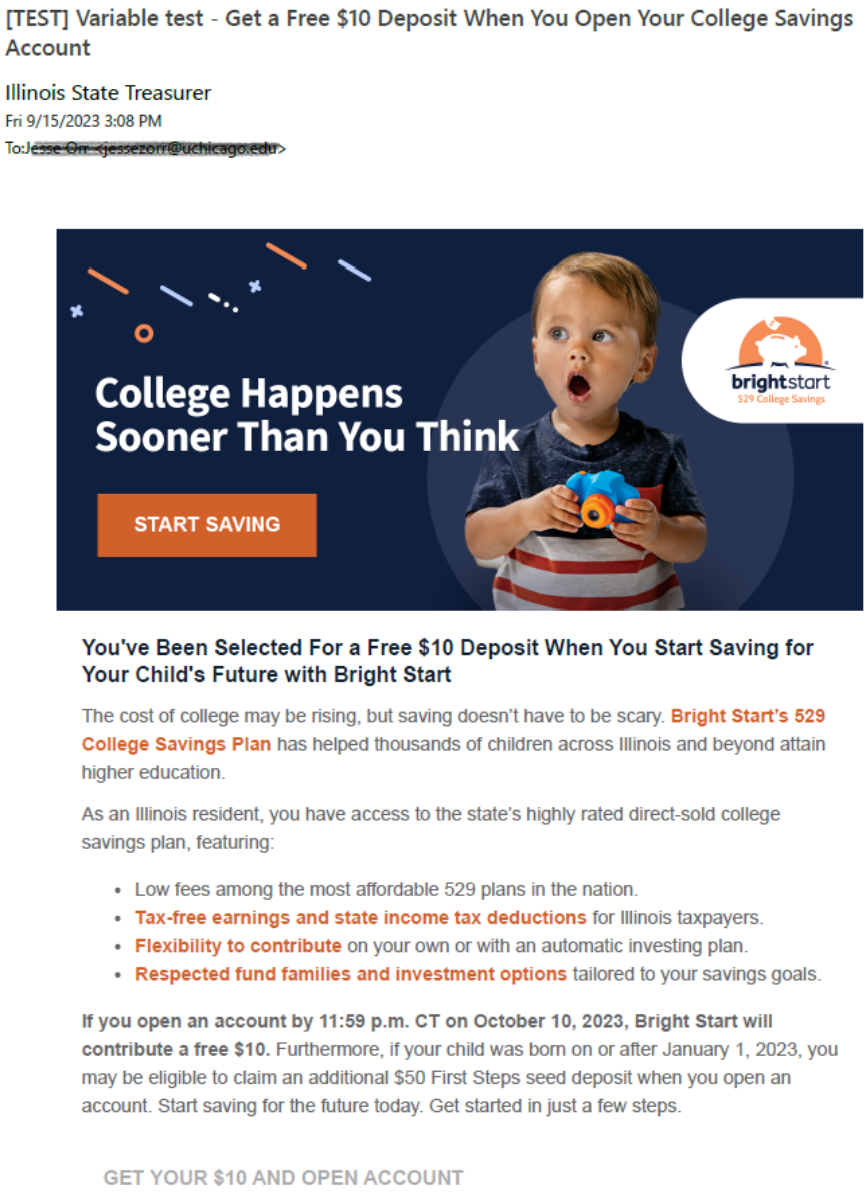
7.1 Field experiment material

Figure B.1: Email used in the field experiment (Control)



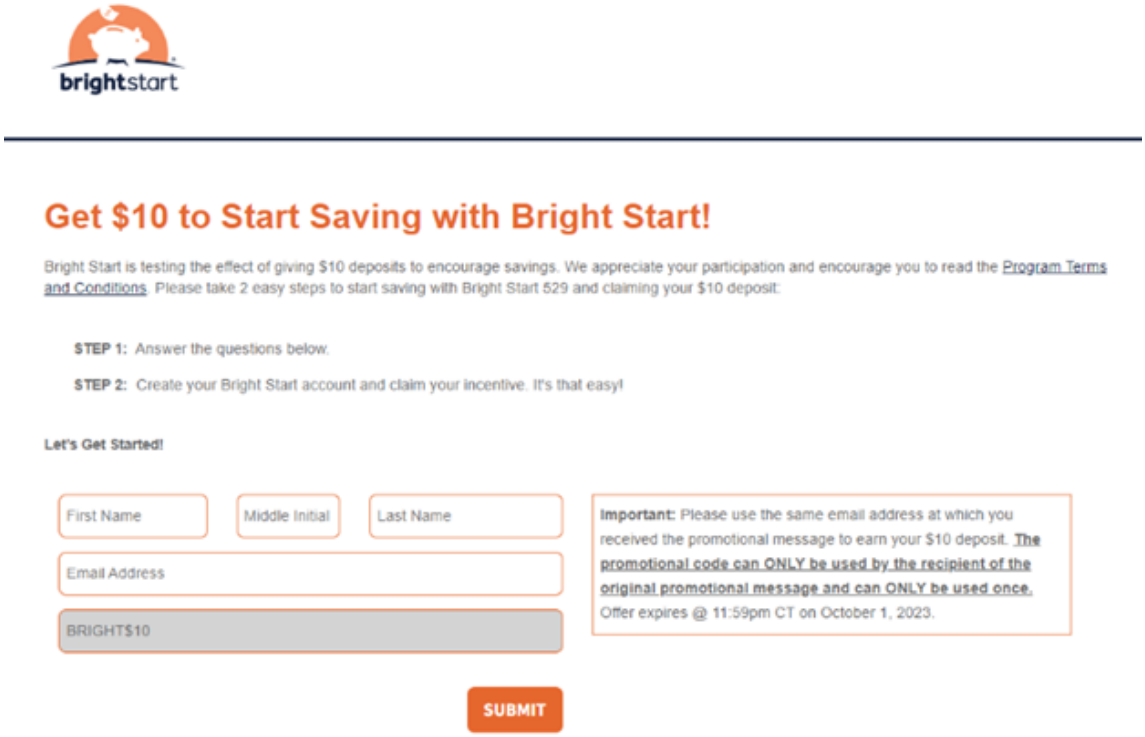
Notes: The figure shows the subject line and content of the marketing email sent to the randomized control group.

Figure B.2: Email used in the field experiment (Treatment)



Notes: The figure shows the subject line and content of the marketing email sent to the randomized treatment groups (the $10 group in this example).

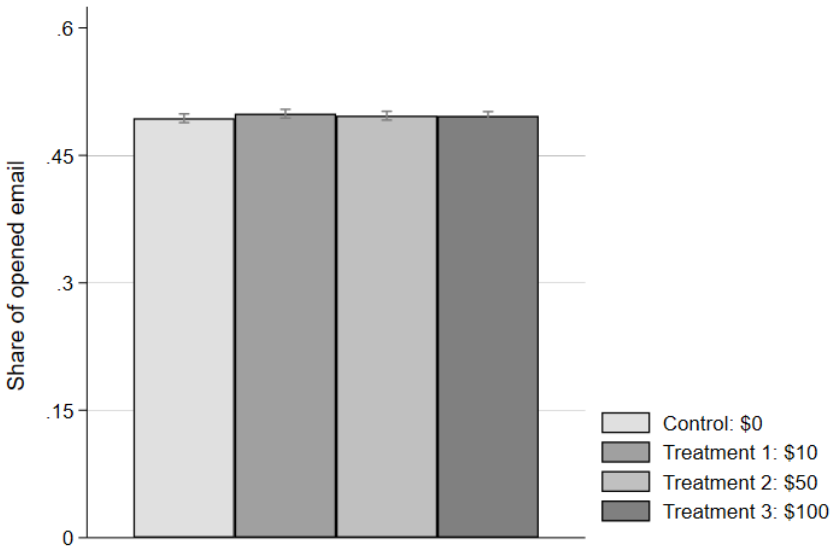
Figure B.3: Landing page after clicking on the link in the invite email



Notes: The figure shows the landing page if email recipients clicked on the link provided in the invite email.

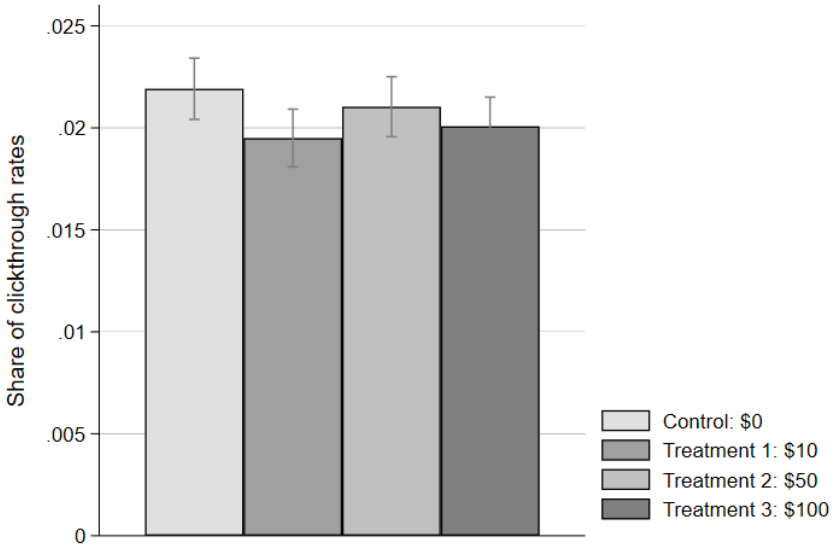
7.2 Field experiment results

Figure B.4: Email opening rates across groups



Notes: The figure shows the average email opening rates across randomized groups, with 95% confidence intervals.

Figure B.5: Click-through rates across groups



Notes: The figure shows the average click-through rate on the link provided in the invite email across randomized groups, with 95% confidence intervals.

Table B.1: Treatment effects in full-scale sample

Treatment	Opened the email as % of email recipients	Clicked on link as % of email recipients	Opened account as % of email recipients	Opened account as % of email openers	Opened account as % of voucher claimed
$0	48.65%	2.16%	0.000%	0.000%	0.00%
$10	49.16%	1.92%	0.003%	0.007%	9.09%
$50	48.90%	2.07%	0.030%	0.074%	40.74%
$100	49.00%	1.98%	0.024%	0.061%	39.13%

Notes: The table shows the outcomes of the full-scale trial.

7.3 Policymakers' survey

7.3.1 Recruitment email

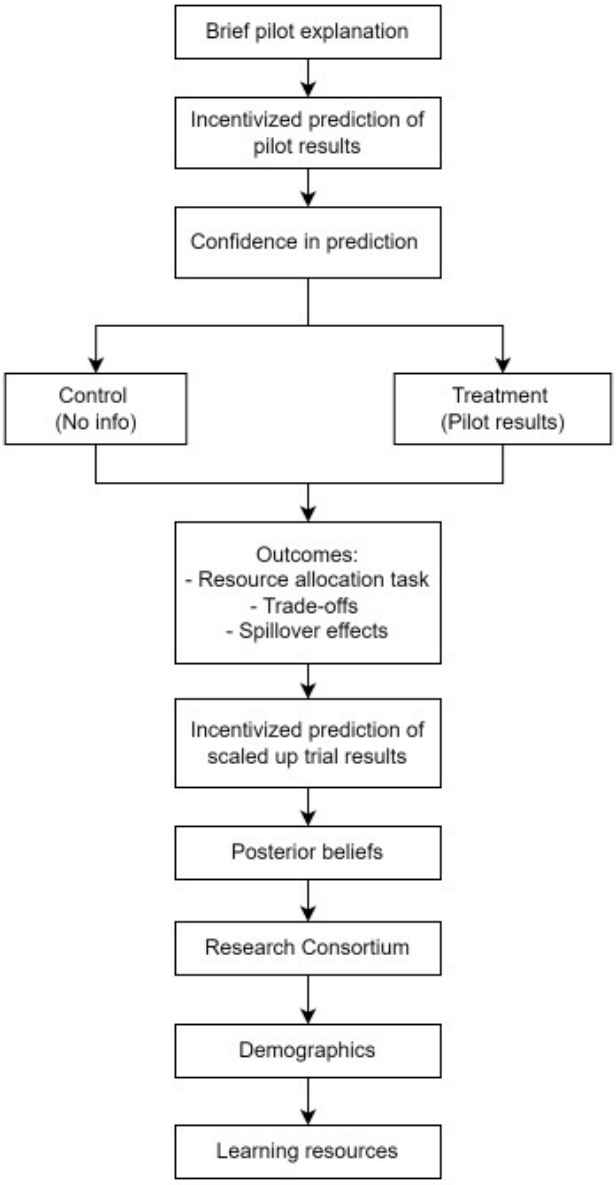
Subject line: Help inform NAST adoption of evidence-based policies and programs

Dear [*name*],
We would like to invite you to take part in a short anonymous survey led by researchers at the University of Chicago. The objective of the survey is to help inform NAST members' adoption of evidence-based decisions to improve program outcomes. We would love to receive your response by Thursday this week. Click here to start the survey: [*Link*]

- The survey should take about 10 minutes to complete and you will have the opportunity to earn up to $50 in Amazon gift cards that you can spend however you want
- You will have the opportunity to learn about an evaluation project to increase the take up rate of a savings program

Thank you for your time. We look forward to receiving your response!
Best,
NAST Conference Organization Committee

Figure B.6: Policymakers' survey design



Notes: The figure shows the flow of the policymakers's survey.

Figure B.7: Survey - trial info page

You've been selected for a free $10 deposit when you start saving for your child's future with a College Savings Account

The cost of college may be rising, but saving doesn't have to be scary. Our **529 College Savings Plan** has helped thousands of children across the state and beyond attain higher education.

As a resident of this state, you have access to the state's highly rated direct-sold college savings plan, featuring:

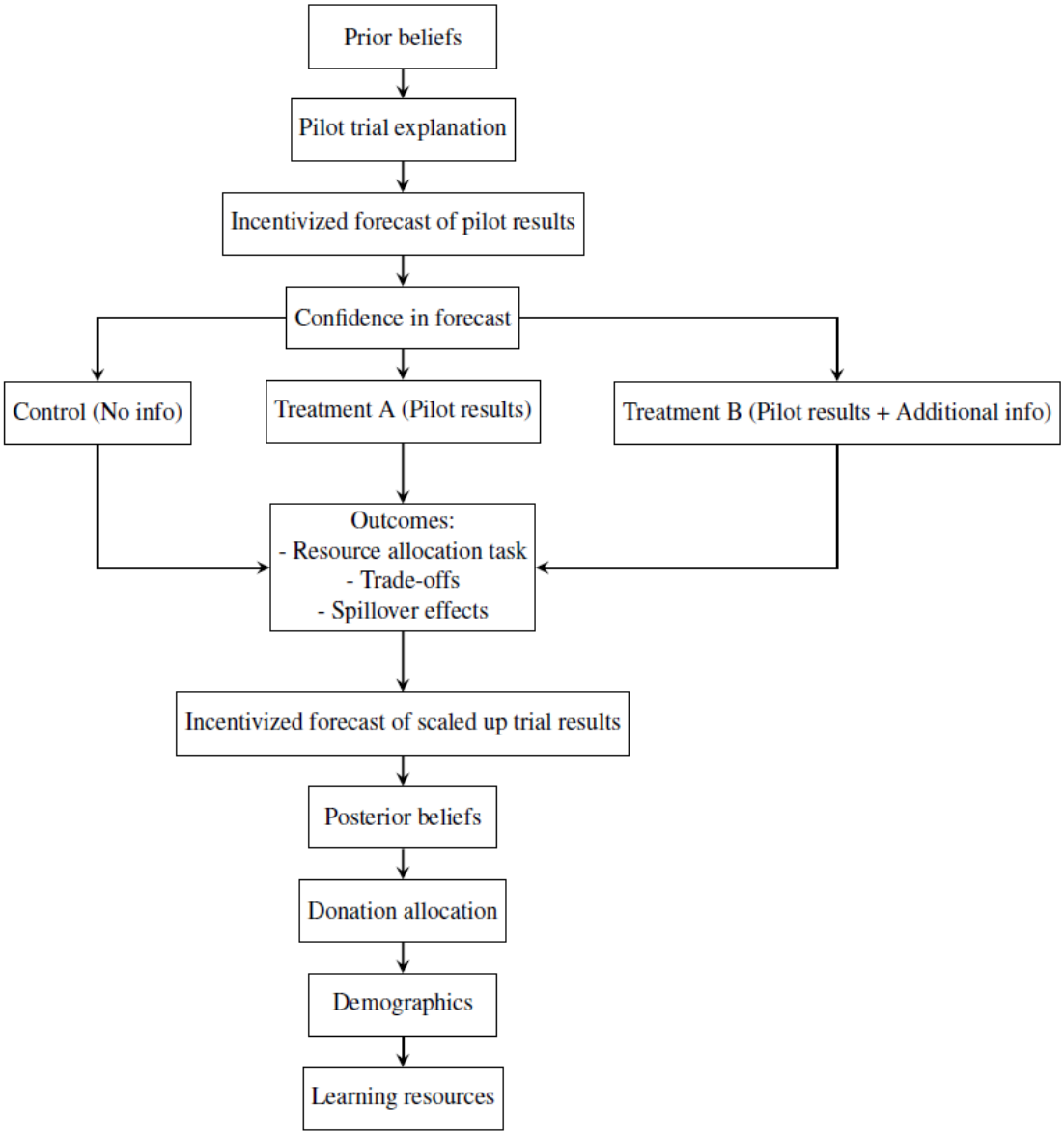
- Low fees among the most affordable 529 plans in the nation.
- **Tax-free earnings and state income tax deductions** for Illinois taxpayers.
- **Flexibility to contribute** on your own or with an automatic investing plan.
- **Respected fund families and investment options** tailored to your savings goals.

If you open an account by 11:59pm on October 10, 2023, our College Savings Plan will contribute a free $10. Furthermore, if your child was born on or after January 1, 2023, you may be eligible to claim an additional $50 seed deposit when you open an account. Start saving for the future today. Get started in just a few steps

Notes: The figure shows the trial info page survey respondents saw

7.4 General public's survey

Figure B.8: Policymakers' survey design



Notes: The figure shows the flow of the general public's survey.

Table B.2: Balance table of the policymakers' survey experiment

	(1)	(2)	(3)
	Control	Treatment	p-values
Age 18-34	0.194	0.254	0.401
	(0.399)	(0.438)	
Age 35-54	0.458	0.577	0.156
	(0.502)	(0.497)	
Age 55+	0.347	0.169	0.015**
	(0.479)	(0.377)	
Confident in pilot priors	0.569	0.549	0.810
	(0.499)	(0.501)	
Postgraduate degree or higher	0.403	0.437	0.684
	(0.494)	(0.499)	
Female	0.472	0.451	0.798
	(0.503)	(0.501)	
Resp's org. used RCTs before	0.278	0.254	0.745
	(0.451)	(0.438)	
Resp's org. uses seeding incentives for takeup	0.528	0.563	0.672
	(0.503)	(0.499)	
Resp. works on 529 plans	0.569	0.521	0.565
	(0.499)	(0.503)	
Has had current job for at least 3 years	0.639	0.634	0.950
	(0.484)	(0.485)	
Expects to keep current job for 3+ years	0.667	0.704	0.631
	(0.475)	(0.460)	
N	72	71	

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.3: Balance table of the general public's survey experiment

				p-values		
	(1)	(2)	(3)	(4)	(5)	(6)
	Control	Treatment A	Treatment B	C vs TA	C vs TB	TA vs TB
Heard of the program before	0.520	0.522	0.578	0.946	0.101	0.115
	(0.500)	(0.500)	(0.495)			
HS or less	0.515	0.488	0.447	0.438	0.056*	0.254
	(0.500)	(0.500)	(0.498)			
College	0.333	0.343	0.369	0.747	0.276	0.442
	(0.472)	(0.475)	(0.483)			
Postgrad	0.153	0.169	0.183	0.522	0.243	0.597
	(0.360)	(0.375)	(0.387)			
Female	0.530	0.485	0.485	0.204	0.203	0.997
	(0.500)	(0.500)	(0.500)			
<$50K HH income	0.417	0.410	0.344	0.840	0.033**	0.053*
	(0.494)	(0.493)	(0.476)			
$50-100K HH income	0.338	0.308	0.354	0.380	0.619	0.169
	(0.473)	(0.462)	(0.479)			
>$100K HH income	0.245	0.281	0.302	0.246	0.073*	0.526
	(0.431)	(0.450)	(0.459)			
Age 18-34	0.488	0.460	0.467	0.439	0.569	0.840
	(0.500)	(0.499)	(0.500)			
Age 35-54	0.390	0.405	0.394	0.655	0.897	0.751
	(0.488)	(0.492)	(0.489)			
Age 55+	0.123	0.134	0.138	0.617	0.511	0.874
	(0.328)	(0.341)	(0.346)			
Republican	0.320	0.291	0.309	0.374	0.739	0.579
	(0.467)	(0.455)	(0.463)			
Democrat	0.338	0.371	0.352	0.327	0.672	0.579
	(0.473)	(0.484)	(0.478)			
N	400	402	398			

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.4: Correlates of forecast treatment effects in the pilot trial among policymakers

Exp. highest take-up rate (Pilot)	(1)	(2)	(3)
Confident in pilot priors	-0.0350	-0.0328	-0.0334
	(0.0379)	(0.0391)	(0.0396)
Educ: Postgrad		0.0233	0.0229
		(0.0411)	(0.0414)
Female		0.0345	0.0341
		(0.0418)	(0.0421)
Resp's org. used RCTs before		-0.00247	-0.00282
		(0.0452)	(0.0455)
Resp's org. uses seeding incentives for takeup		0.0371	0.0374
		(0.0493)	(0.0496)
Resp. works on 529 plans		-0.0283	-0.0281
		(0.0492)	(0.0494)
Has had current job for at least 3 years		0.0419	0.0412
		(0.0484)	(0.0489)
Expects to keep current job for 3+ years		0.0373	0.0374
		(0.0649)	(0.0651)
Age 35-54		-0.0764	-0.0763
		(0.0673)	(0.0676)
Age 55+		-0.135*	-0.135*
		(0.0733)	(0.0736)
Seconds spent reading trial info			-1.43e-06
			(1.32e-05)
Constant	0.269***	0.255***	0.256***
	(0.0284)	(0.0916)	(0.0927)
N	143	143	143
R^2	0.006	0.045	0.045

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.5: Correlates of forecast treatment effects in the pilot trial among the general public

Exp. highest take-up rate (Pilot)	(1)	(2)	(3)	(4)
Confident in pilot priors	0.0557***	0.0514***	0.0479**	0.0497**
	(0.0190)	(0.0193)	(0.0193)	(0.0195)
Heard of the program before		-0.0425**	-0.0411**	-0.0417**
		(0.0171)	(0.0170)	(0.0172)
Educ: College		-0.0237	-0.0226	-0.0237
		(0.0184)	(0.0184)	(0.0184)
Educ: Postgrad		-0.0139	-0.0129	-0.0141
		(0.0246)	(0.0246)	(0.0246)
Female		0.0279*	0.0257	0.0252
		(0.0161)	(0.0161)	(0.0161)
Age 35-54		-0.0313*	-0.0343**	-0.0320*
		(0.0172)	(0.0172)	(0.0175)
Age 55+		-0.0152	-0.0213	-0.0192
		(0.0251)	(0.0252)	(0.0256)
$50-100K HH income		0.0233	0.0251	0.0272
		(0.0190)	(0.0190)	(0.0191)
>$100K HH income		0.00541	0.00733	0.00951
		(0.0215)	(0.0215)	(0.0217)
Believes the program helps		0.0215***	0.0207***	0.0216***
		(0.00731)	(0.00732)	(0.00739)
Seconds spent reading trial info			0.000241*	0.000227*
			(0.000123)	(0.000124)
Republican				0.0103
				(0.0253)
Democrat				0.0294
				(0.0236)
State Gov is same political party				-0.0258
				(0.0212)
Constant	0.368***	0.276***	0.263***	0.255***
	(0.00899)	(0.0436)	(0.0439)	(0.0448)
N	1,200	1,200	1,200	1,200
R^2	0.007	0.028	0.031	0.034

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.6: Trust Spillovers

Posterior trust beliefs in own state Governor	(1) Competence	(2) Competence	(3) Benevolence	(4) Benevolence	(5) Integrity	(6) Integrity
Treatment A	-0.210***	-0.206***	-0.0736	-0.0701	-0.144**	-0.130**
	(0.0669)	(0.0663)	(0.0664)	(0.0651)	(0.0633)	(0.0617)
Treatment B	-0.101	-0.0969	0.0856	0.0966	-0.119*	-0.103
	(0.0619)	(0.0608)	(0.0598)	(0.0590)	(0.0646)	(0.0637)
Priors	0.836***	0.798***	0.828***	0.798***	0.846***	0.811***
	(0.0155)	(0.0173)	(0.0155)	(0.0172)	(0.0147)	(0.0170)
Confident in pilot priors		-0.0104		0.0564		0.0864
		(0.0663)		(0.0690)		(0.0664)
Exp. highest take-up rate (Pilot)		0.0612		-0.167*		0.0255
		(0.0949)		(0.0935)		(0.0950)
Heard of the program before		0.0605		-0.0680		0.0403
		(0.0559)		(0.0547)		(0.0559)
Believes the program helps		0.142***		0.0835***		0.0847***
		(0.0261)		(0.0247)		(0.0270)
Republican		-0.112		-0.124		-0.122
		(0.0852)		(0.0882)		(0.0822)
Democrat		-0.0294		0.0671		-0.104
		(0.0803)		(0.0750)		(0.0715)
State Gov is same political party		0.238***		0.256***		0.322***
		(0.0784)		(0.0750)		(0.0675)
Constant	0.644***	0.0389	0.595***	0.236	0.637***	0.0772
	(0.0804)	(0.155)	(0.0774)	(0.156)	(0.0682)	(0.154)
Controls	No	Yes	No	Yes	No	Yes
N	1,200	1,200	1,200	1,200	1,200	1,200
R^2	0.694	0.711	0.717	0.729	0.727	0.739
Test: Treat A = Treat B	0.1048	0.0981	0.0123	0.0079	0.6856	0.6613

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH income and number of seconds spent reading the trial info page.

Table B.7: General Public' Support for RCT - Validation

Donation share to RCT charity	(1)	(2)	(3)	(4)
Treatment A	0.00300	0.00326	0.000101	0.000151
	(0.0161)	(0.0161)	(0.0160)	(0.0160)
Treatment B	-0.000978	0.000399	-0.00179	-0.00168
	(0.0161)	(0.0161)	(0.0161)	(0.0161)
Confident in pilot priors		0.0247	0.00860	0.00940
		(0.0158)	(0.0160)	(0.0161)
Exp. highest take-up rate (Pilot)		-0.0295	-0.0305	-0.0298
		(0.0240)	(0.0240)	(0.0241)
Heard of the program before			-0.0160	-0.0163
			(0.0142)	(0.0142)
Believes the program helps			0.0125**	0.0126**
			(0.00609)	(0.00610)
Constant	0.464***	0.469***	0.445***	0.448***
	(0.0114)	(0.0148)	(0.0380)	(0.0383)
Controls	No	No	Yes	Yes
N	1,200	1,200	1,200	1,200
R^2	0.000	0.003	0.029	0.029

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH incomes. Model (4) also includes number of seconds spent reading the trial info page.

Table B.8: Multiple Hypothesis Testing on Trust Outcomes

			p-values				
			Unadjusted	Multiplicity Adjusted			
Outcome	Treatment	Coeff.	Remark 3.2	Theorem 3.1	Remark3_8	Bonferroni	Holm
Δ competence	A	-.2217119	0.0023333	0.0116667	0.0116667	0.014	0.014
Δ competence	B	-.0361016	0.5873333	0.5873333	0.5873333	1	0.5873333
Δ benevolence	A	-.1164769	.1333333	.2486667	.2486667	.8	0.2666667
Δ benevolence	B	.1245553	.0756667	.2026667	.2026667	.454	0.227
Δ integrity	A	-.1591309	.019	.085	.085	.114	0.095
Δ integrity	B	-.1519337	.0373333	.1356667	.1356667	.224	0.1493333

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH incomes. Model (4) also includes number of seconds spent reading the trial info page.

Table B.9: Framing effects

	(1)	(2)	(3)	(4)	(5)	(6)
Posterior trust beliefs in STO	Δ competence	Δ competence	Δ benevolence	Δ benevolence	Δ integrity	Δ integrity
Treatment A	-0.452	-0.271	-0.671	-0.271	0.473	0.514
	(0.385)	(0.560)	(0.610)	(0.560)	(0.444)	(0.622)
Treatment B	-0.655*	-0.459	-0.526	-0.459	-0.00899	0.516
	(0.362)	(0.504)	(0.648)	(0.504)	(0.419)	(0.474)
Survey version dummy	-0.164	0.108	-0.214	0.108	0.0458	0.541
	(0.320)	(0.361)	(0.419)	(0.361)	(0.345)	(0.655)
Treat. A * Survey version dummy		-0.382		-0.382		-0.283
		(0.821)		(0.821)		(0.881)
Treat. B * Survey version dummy		-0.442		-0.442		-1.373
		(0.740)		(0.740)		(0.908)
Confident in pilot priors	-0.278	-0.289	-0.0579	-0.289	0.371	0.385
	(0.339)	(0.358)	(0.502)	(0.358)	(0.322)	(0.331)
Exp. highest take-up rate (Pilot)	0.0332	0.0736	-0.234	0.0736	0.504	0.576
	(0.445)	(0.470)	(0.697)	(0.470)	(0.505)	(0.504)
Heard of the program before	-0.194	-0.260	-0.1000	-0.260	0.670**	0.450
	(0.286)	(0.326)	(0.391)	(0.326)	(0.313)	(0.360)
Believes the program helps	0.187	0.175	0.157	0.175	-0.00527	-0.0466
	(0.168)	(0.164)	(0.162)	(0.164)	(0.194)	(0.160)
Republican	0.299	0.310	0.431	0.310	1.313**	1.250***
	(0.638)	(0.683)	(0.817)	(0.683)	(0.496)	(0.461)
Democrat	0.904	0.939	0.426	0.939	0.629	0.681
	(0.882)	(0.931)	(1.118)	(0.931)	(0.737)	(0.676)
State Gov is same political party	-0.404	-0.352	0.130	-0.352	-0.711	-0.622
	(0.788)	(0.838)	(1.096)	(0.838)	(0.610)	(0.603)
Constant	-0.396	-0.459	-0.527	-0.459	-0.457	-0.313
	(1.002)	(0.977)	(0.980)	(0.977)	(1.155)	(0.988)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	45	45	45	45	45	45
R^2	0.338	0.347	0.443	0.347	0.574	0.627

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Control variables include: College/Postgrad Educ, Female, Age 35-54 or 55+, $50-100K or >$100K HH incomes. Model (4) also includes number of seconds spent reading the trial info page.