Understanding and Increasing Policymakers' Sensitivity to Program Impact*

Mattie Toma† Elizabeth Bell

October 10, 2022

Click here for most recent version

Abstract

Policymakers routinely make high-stakes funding decisions. In two experiments with policymakers in the U.S. government and the general public, we find that valuations of programs are inelastic with respect to program impact. We design and test two decision aids, one which presents programs side-by-side and another which translates multiple features of impact into an aggregate metric. The decision aids increase elasticity by 0.20 on a base of 0.33 among policymakers and by 0.21 on a base of 0.21 among the general public. We provide evidence that the difficulty of assessing complex inputs can help explain the inelasticity of program valuations.

*We thank Alec Brandon, Elizabeth Engle, Ben Enke, Christine Exley, Clayton Featherstone, Zach Freitas-Groff, Justin Holz, Woojin Kim, David Laibson, Steven Levitt, Matthew Rabin, Gautam Rao, Lisa Robinson, Todd Rogers, Rohen Shah, Neil Stewart, Cass Sunstein, Michael Thaler, Eva Vivalt, as well as numerous seminar participants for helpful feedback. We would also like to thank the team at the Office of Evaluation Sciences in the U.S. General Services Administration, particularly Kelly Bidwell, Julia Brown, Russ Burnett, and Elana Safran for their unflagging support of this project. The evidence teams at the U.S. Department of Education including Matt Soldner, the U.S. Department of Justice, the U.S. Department of Health and Human Services, the U.S Department of the Treasury, and the U.S. Agency for International Development, as well as all participating policymakers, generously contributed their time and insights. Collin Cox, John-Henry Pezzuto, and Abigail Tikhtman provided excellent research assistance. We gratefully acknowledge funding and support from the Mind Brain Behavior graduate award and the Centre for Effective Altruism. The experiment was pre-registered with a pre-analysis plan on the AEA registry, number AEARCTR-0007659. We received IRB approval from Harvard University, protocols IRB21-0002 and IRB21-0989.

†Toma: Warwick Business School, Office of Evaluation Sciences, and Global Priorities Institute, University of Oxford (Mattie.Toma@wbs.ac.uk); Bell: Florida State University and Office of Evaluation Sciences (ebell3@fsu.edu)

1 Introduction

In recent years, both researchers and practitioners have allocated substantial attention and resources to generating evidence about the efficacy of government programs and interventions. The Foundations for Evidence-based Policymaking Act of 2018 grew out of this broader focus, setting the stage for an ambitious agenda on the advancement of evidence production and use by the U.S. government (H.R.4174, 2019). However, for evidence to effectively impact decisions about which programs to implement, policymakers must be well-equipped to use scientific findings to inform their assessments of program value.

Large and varied literatures document the cognitive difficulty and uncertainty non-experts face when assessing the value of a broad range of goods and services.[1] Given that evidence utilization in program funding decisions requires both a theory of how to translate information about varied and complex program features into an estimate of program value as well as the bandwidth to execute this translation, such difficulties may extend to expert policymakers as well. Indeed, when we surveyed high-ranking federal employees and asked them to indicate the practical barriers to evidence utilization from a list of ten factors such as a lack of time or funding, the most common responses were "uncertainty about how to turn evidence into action" and "difficulty interpreting the implications of evidence-based recommendations" (Appendix Figure A.1).

Taking this hypothesized role of the cognitive difficulty involved in program funding decisions as a point of departure, we leverage an experiment among high-ranking federal employees, recruited across 22 U.S. government agencies, who interact with program-relevant evidence as part of their job. Our paper makes three main contributions. First, we find that policymakers' assessments of the value of a policy program are inelastic with respect to impact; that is, their valuations don't update one-to-one with a change in program impact. Second, we test two decision aids that aim to simplify the decision problems and show that they increase policymakers' responsiveness, or sensitivity, to evidence-based information about a program. Third, using the documented treatment effects as well as correlational data, we provide evidence indicating that the cognitive difficulty of mapping scientific findings to an assessment of program value plays an important role in driving the observed inelasticity. Together, our results suggest that program valuations made by expert policymakers are subject to bias, and that simple and portable decision aids can address this bias.

To estimate the degree to which policymakers update their assessments of a program's value in response to impact-relevant information, policymakers in our experiment see de-

[1] This is shown in the behavioral economics research on bounded rationality (Alós-Ferrer et al., 2021; Benjamin, 2019; Simon, 1955) as well as a long literature on stated preferences (Dickert et al., 2015; Kahneman and Knetsch, 1992).

scriptions of six randomly-assigned hypothetical policy programs, typically tailored to the type of work they do in government. Each program description includes details on three features of evidence relevant to impact, which are randomized across respondents: "scope" (number of people reached), "outcome type" (whether the program affects the downstream outcomes ultimately of interest or intermediate outcomes), and "persistence" of effects (how long the program effects last). In our analysis we use these three parameters to calculate a quantitative measure that captures the person-years of program impact. After seeing each program and its impact-relevant features, respondents provide assessments of the value of the program by indicating the maximum cost at which they would be willing to support funding the program out of their department's budget. We first estimate the elasticity of responses with respect to impact in the control condition—when assessments are made without a decision aid—by comparing assessments of the value of a program to our quantitative measure of program impact. If program assessments scaled one-to-one with a change in impact, we would estimate an elasticity of 1. Instead, we find that policymakers' assessments are markedly inelastic with respect to impact.

In part, this inelasticity may simply reflect preferences. Policymakers' true value of a program may not be linear in impact—for instance, due to career incentives policymakers may derive a fixed value from launching a new program irrespective of, say, the precise number of beneficiaries. Structural barriers including budget constraints and organizational inertia may also hinder evidence-based decision-making (DellaVigna et al., 2022; Lugo-Gil et al., 2019; Natow, 2020). Alternatively, behavioral biases may play a role in distorting how policymakers interpret information about programs and policies.[2] Of particular relevance, a long literature on contingent valuation shows that individuals engage in "scope neglect": they are poorly attuned to the number of people affected by a program when making assessments of how much the program is worth (Dickert et al., 2015; Kahneman and Knetsch, 1992). In this paper, we provide both causal and correlational evidence suggesting that policymakers—experts in these types of decisions—place less weight on impact-relevant, evidence-based features of programs due to the cognitive complexity of the decision environment.

To provide causal evidence for a role of bounded rationality as well as to identify practical policy solutions, we test two decision aids that increase the elasticity of assessments of program value with respect to impact by 60% ($p = 0.004$). By showing that policymakers' assessments of programs are malleable, these interventions indicate that the observed inelasticity is not driven (solely) by policymakers' preferences. Moreover, because the decision

[2]Confirmation bias, motivated reasoning, status quo bias, the effects of framing on risk aversion, variance neglect, and overconfidence have all been either hypothesized or shown to factor into policy-relevant decisions (Banuri et al., 2019; Christensen and Moynihan, 2020; Hjort et al., 2021; Mayar et al., 2021; Vivalt and Coville, 2021; World Bank Group, 2015).

aids were designed to simplify the mapping between the impact-relevant information policymakers receive and their assessments of the program's dollar value, the efficacy of the aids suggests that the cognitive difficulty of assessing a program's value is an important barrier.

The first decision aid we employ to increase policymakers' sensitivity to impact applies insights from psychology and marketing indicating that when options are presented simultaneously rather than sequentially, decisions are more consistent and people put more weight on difficult-to-evaluate attributes (Hsee, 1996; Hsee et al., 1999; Bohnet et al., 2016). In our experiment we randomly vary, within-subject, whether information about one policy program is presented in isolation on a decision screen or if two similar programs appear side-by-side. Consistent with our hypothesis, when respondents see two similar programs together, their observed sensitivity to impact is 79% greater ($p = 0.001$); this "Side-by-Side" decision aid increases the elasticity of assessments of program value with respect to impact by 0.26 on a base of 0.33.

The second decision aid involves randomly presenting respondents with an "Impact Calculator" that translates total program costs into an annual cost per person impacted. This calculator does not add any new information: the numbers required to calculate the aggregate metric for impact are all clearly available in the Control condition as well. However, we hypothesize that the provision of our Impact Calculator will ease the cognitive burden of assessing impact even for experts and, in turn, will facilitate increased sensitivity when assessing program values.[3] Indeed, we see that when policymakers are presented with an Impact Calculator, the elasticity of assessments of program value with respect to impact increases by 0.20 on the base of 0.33 ($p = 0.024$). This reflects a 60% increase compared to the sensitivity observed in the Control condition.

The substantial effects of the two interventions—which simplify the decision problem in different ways—point to the role of bounded rationality in reducing sensitivity to impact. Additional correlational evidence also supports this mechanism. For one, using the tool developed in Enke and Graeber (2021) to measure cognitive uncertainty, we ask respondents to self-report the degree of certainty they experienced when making decisions on a scale from 0 to 100. We see that certainty in one's assessments is positively correlated with increased sensitivity in the Control condition ($p = 0.015$), suggesting that those who experience more difficulty or confusion when translating the impact-relevant information into program assessments are also those whose assessments are less responsive to information about impact. Furthermore, both the Side-by-Side presentation and the Impact Calculator make people

[3]This hypothesis is consistent with lab experiments among the general public, where Boyce-Jacino et al. (2021) and Saiewitz and Piercey (2019) examine people's ability to assess federal budgetary expenditures when presented in total versus per capita terms and find that people are better able to distinguish between numbers once the dollar amounts have been converted into more digestible units.

more certain in the assessments they gave. Consistent with the notion that sensitivity to impact is higher when an individual's "theory" of how to map impact-relevant information onto program value is more developed and clearly-defined, we also see a significant positive relationship between sensitivity and a measure of real-world experience with the types of program assessments respondents encounter in the experiment ($p < 0.001$).

We contrast our findings among expert policymakers with the general public by conducting a similar experiment among a representative sample of 500 U.S. citizens through the online platform Prolific. We observe even lower sensitivity to program impact among non-experts, with the elasticity of assessments of program value with respect to impact equal to 0.21 among the general public compared to 0.33 among policymakers. We also observe large treatment effects in this population: The Impact Calculator and Side-by-Side decision aids increase elasticity by 0.21 on this base of 0.21 ($p < 0.001$). To learn whether the effect of the decision aids are additive, we test a third decision aid that combines these two interventions and find that treatment effects roughly double when the decision aids are presented together. This arm of the experiment also allows us to shed additional light on mechanisms by introducing a module at the end of the experiment to capture participants' numeracy. Consistent with bounded rationality as a key driver of low sensitivity to program impact, we observe a strong correlation between numeracy and sensitivity. Finally, we replicate the central findings in an incentivized version of the experiment in which respondents receive payments based on their predictions of the modal program assessments provided by other participants.

This paper contributes to an emerging literature on policymakers' ability to interpret and utilize evidence-based information in programmatic decision-making. Under certain conditions, research suggests that evidence can impact program adoption decisions. In Brazil, Hjort et al. (2021) find that mayors demand evidence-based program information, and that access to impact evaluations can affect policy adoption, although the effects are arguably modest. Mehmood et al. (2021) show that an econometrics training program increases demand for evidence among bureaucrats, while Crowley et al. (2021) find that an intervention involving outreach around legislative use of research evidence is effective. There is also work indicating that policymakers may be less attuned to complexities in the content of the evidence. For example, Vivalt et al. (2021) and Nakajima (2021) find that policymakers place relatively more weight on external validity and less on features relevant to internal validity, despite the importance of internal validity for determining the quality of the evidence. Perhaps most concerning is the research in the realm of education policy, which finds that many policymakers lack the skills to critically evaluate the quality of evidence or the norms to understand and incorporate evidence into decision-making (Hill and Briggs, 2020; Bergman et

al., 2020; Moynihan and Lavertu, 2012). To our knowledge, our paper is the first to directly estimate (and seek to increase) *sensitivity to program impact* in a policymaking context.

This paper also adds to a literature in behavioral and experimental economics that seeks to understand how complex decision environments generate various types of under-sensitivities in economically-relevant behavior, including choice under risk, consumer choice, and belief updating (Alós-Ferrer et al., 2021; Benjamin, 2019; Enke and Graeber, 2021; Khaw et al., 2020; Tversky and Kahneman, 1974). This work demonstrates that the cognitive difficulty of processing information can lead individuals to not fully update based on new information, resulting in decisions that are attenuated with respect to the relevant inputs. As noted above, the contingent valuation literature similarly captures an under-responsiveness among the general public to features such as program scope.[4] Our paper adds to this literature on under-sensitivities in complex decision environments by contributing some of the first field evidence that directly links these theoretical concepts and ideas typically tested in the lab to policy-relevant decision contexts among a sample of experts.

The rest of the paper proceeds as follows: Section 2 develops the conceptual framework for the application of bounded rationality in program adoption decisions. Section 3 outlines the design of the lab-in-the-field experiment among policymakers. Section 4 presents the elasticity observed at baseline in the experiment as well as the efficacy of the decision aids and potential mechanisms. Section 5 details the design and results of the complementary experiment among the general public. Section 6 discusses theoretical and practical implications of these findings and highlights promising opportunities for future research.

2 Conceptual Framework

To understand the factors that might limit policymakers' utilization of evidence-based information about program impact when making funding decisions, it is useful to conceptualize the policymaker's decision problem. To fix ideas, consider a policymaker who is choosing whether to fund a program at cost c. She values both program impact, m, and other program factors x independent of m, for instance the program's political appeal and visibility or the characteristics of the population affected by the program. If she could fully conceptualize how to incorporate information about program impact, her utility would be $u(m, x, c) = wm + (1 - w)x - c$, where w represents the weight she places on m versus x. She would fund the program if and only if $wm + (1 - w)x \geq c$.

[4]There exists a strand of literature in psychology testing interventions to improve responsiveness to the scope of a stimulus in lab environments, although the results have not been consistently replicated (Evangelidis and den Bergh, 2013; Hsee and Rottenstreich, 2004; Small et al., 2007).

However, the difficulty of the decision problem may play a role in limiting the policymaker's *sensitivity* to program impact. Intuitively, it may be difficult for the policymaker to know what information about program impact means in terms of how she ought to update about the total dollar value of the program. In other words, limited sensitivity may stem from a lack of a clear blueprint from which to translate impact-relevant information into concrete assessments of program value. Sensitivity may be further depressed by limited bandwidth due to time, attention, or cognitive capacity constraints.

More formally, we posit that the policymaker interprets impact, m, as $\hat{m} \equiv \lambda m + (1-\lambda)p$, where p is her prior on the value of the program and λ is her sensitivity to impact. In this case, she will value the program at $w(\lambda m + (1-\lambda)p) + (1-w)x - c$. That is, in difficult decision problems policymakers will reduce their sensitivity to impact from w to λw.

This framework is a stylized version of models of cognitive imprecision as they are increasingly found in the economics and psychology literatures (e.g. Enke and Graeber (2021), Gabaix (2019), and Khaw et al. (2020)). Importantly, we do not assume that the policymaker knows a "correct" formula for translating impact into assessments of program value and is simply implementing this noisily. Instead, the policymaker starts with a default or anchor, p, and then adjusts in the direction of the evidence to a degree consistent with their sensitivity, λ. It is also worth noting that the intent of this paper is not to pin down the particular micro-foundations underlying this effect. For instance, while a model of cognitive imprecision akin to Enke and Graeber (2021) or Khaw et al. (2020) would be consistent with the framework we've outlined, heuristics akin to Tversky and Kahneman's 1974 model of adjustment and anchoring would produce a similar pattern. The common theme of these mechanisms is that the difficulty of the decision problem limits policymakers' ability to fully incorporate the information they are receiving, leading to attenuated responses driven by lower sensitivity to program impact.

Figure I provides a simplified sketch (in log-log scale) of three possible approaches a decision maker might use to assess program value based on impact, as motivated by the conceptual framework described above. The x-axis reflects a measure of program impact (the person-years of impact) for different programs while the y-axis shows a decision maker's assessment of the dollar value of the programs. The solid line identifies program assessments that scale one-to-one with a change in impact (that is, the case where $w, \lambda = 1$), which represents a natural benchmark to compare alternative decision rules.

The dashed black line, meanwhile, identifies a decision maker whose assessments of program value are inelastic with respect to impact. Preferences for program factors unrelated to impact, x, may contribute to relatively attenuated sensitivity to program impact. In this experiment we examine whether sensitivity to impact in the policymaking context is

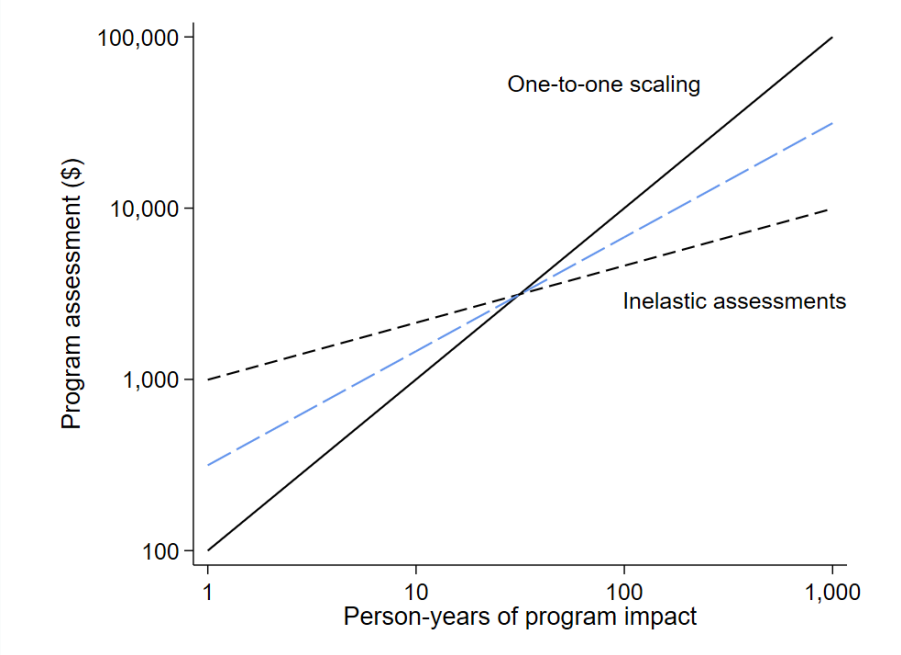


FIGURE I: Simplified Sketch of Sensitivity in Program Adoption Decisions

This figure plots hypothetical program assessments provided by three decision makers. The x-axis shows the person-years of impact as a simple measure of program impact. The y-axis shows the dollar value decision makers assign to the programs, depending on their impact. The policymaker whose assessments of program value scale one-to-one with changes in program impact is reflected by the solid black line. The policymaker with inelastic assessments (represented by the dashed black line) updates in this example by placing equal weight on the impact-relevant information and their prior, resulting in a flatter slope in comparison to the assessments that scale one-to-one. Finally, decision aids that simplify the decision problem increase a policymaker's sensitivity to impact, or weight placed on the impact-relevant information, as reflected by the blue dashed line.

additionally limited due to the complexity of the decisions involved in making program assessments, captured by $\lambda < 1$. In the example in Figure I, the decision maker represented by the dashed black line updates by placing equal weight on the impact-relevant information and their prior, such that lower sensitivity to program impact leads to attenuation in the direction of their prior. Lower sensitivity to impact may also affect the intercept, with the potential to lead, for instance, to under-valuations when program impact is high or over-valuations when impact is low. Finally, under this framework clarifying the decision problem will increase the policymaker's sensitivity to program impact, λ, which is reflected in the blue dashed line in Figure I.

This paper tests three hypotheses based on this conceptual framework:

Hypothesis 1: Assessments of program value are inelastic with respect to impact-relevant information.

In the experiment we test whether policymakers' assessments of program value are inelastic with respect to evidence about a program's scope, outcome type, and persistence.

Hypothesis 2: Decision aids that clarify the mapping between program impact and value increase sensitivity to impact.

We test the efficacy of two decision aids aimed at simplifying the process of interpreting and assessing impact-relevant information about a program. A positive effect of these decision aids on sensitivity to impact indicates that individuals were not optimizing at baseline; that is, if the framing of impact-relevant information matters, then we can infer that assessments at baseline are not a direct indicator of preferences, and indeed that it is likely that $\lambda < 1$. Figure I depicts the hypothesized mechanistic role of the decision aids in increasing the elasticity of assessments of program value with respect to impact via the blue dashed line.

Hypothesis 3: Proxies for policymakers' ability to assess impact are correlated with sensitivity to impact.

We predict a positive relationship between sensitivity and several plausible proxies for a more clear-cut understanding of how to incorporate information about program impact. These proxies include certainty in one's responses, experience with programs and evaluations, and numeracy. For instance, according to the conceptual framework outlined above, when the policymaker finds it more difficult to map information onto assessments of program value, we expect that she will report less certainty in her assessments and will also update her assessments less in response to program impact.

3 Experimental design

3.1 Data

The lab-in-the-field experiment is conducted through the Office of Evaluation Sciences (OES), an office in the United States General Services Administration (GSA) that designs and evaluates evidence-based programs and program changes informed by the social and behavioral sciences.

3.1.1 Population of policymakers

Study participants are 191 employees across 22 of 24 U.S. federal government agencies whose roles involve developing, interpreting, and/or making adoption decisions based on program-relevant evidence. As is expected given these roles, participants in the experiment

hold high-ranking positions in government. The modal respondent is a Grade GS-14 employee, where the General Schedule payscale in the federal government ranges from Grade GS-1 to Grade GS-15. As shown in Appendix Table A.1, 28% of respondents work at the U.S. Department of Education (Ed); 23% work at the U.S. Department of Health and Human Services (HHS); 7% work at the U.S. General Services Administration (GSA); 5% work at the United States Agency for International Development (USAID); 5% work at the U.S. Department of Justice (DOJ); and the remaining 32% are from one of the 19 other agencies in the U.S. government.

Participants were recruited via one of three approaches, in order of how common each approach was: (1) we presented a high-level overview of our research proposal to an Evaluation Officer or other Evidence Lead within a federal agency, and these leads then worked with us to identify relevant policymakers in their agency; (2) the Office of Evaluation Sciences identified policymakers across government who signed up for or attended regular OES workshops on evidence and analysis; and (3) experiment respondents recommended others in their agency who fit the criteria to participate.[5] Selected individuals were invited to take part in a short online survey hosted by the Office of Evaluation Sciences. An example recruitment message can be found in Appendix Figure E.1. Policymakers were recruited from May to October, 2021. 1,469 policymakers received an invitation to take the survey.[6] Of these, 191 completed the survey, for an overall response rate of 13%.[7]

3.1.2 Evidence use in the U.S. government

Across the federal government in the United States, bureaucrats are being encouraged to infuse evidence in decision-making. Previous government reforms such as the Government Performance and Results Act (GPRA) and the Performance Assessment Rating Tool (PART) sought to advance evidence-based policymaking by increasing the body of evidence generated by the federal bureaucracy on program outcomes and performance (Moynihan and Pandey, 2010; Moynihan and Lavertu, 2012). With the 2021 White House Memorandum on *Restoring Trust in Government Through Scientific Integrity and Evidence-Based Policymaking* and the Office of Management and Budget's guidance elevating program evaluations to "critical agency function[s]", it has become clear that evidence generation is now an integral aim of

[5] We also recruited one additional respondent by advertising the survey in an OES newsletter.

[6] It is possible that additional individuals saw our invitation, for instance in the OES newsletter; this number includes all individuals who either received a personalized invitation or an email from an agency Evidence Lead inviting a list of relevant federal employees to participate.

[7] Consistent with the method posted in our pre-registration, we consider a survey "complete" and include the response in this total once a respondent has answered our main program assessment questions, even if they do not complete all follow-up questions about their background and experience with evidence and evaluation.

policymaking (Executive Office of the President, 2021; Vought, 2021). However, the evidence generated by these reforms has often gone unused (Moynihan and Pandey, 2010; Haskins and Margolis, 2014; Natow, 2020; Moynihan and Lavertu, 2012), and what remains unclear is the degree to which policymakers are equipped to interpret and utilize scientific evidence. The policymakers recruited for our experiment are those directly impacted by and acting on these reforms, and as such our findings regarding the nature and extent of sensitivity to impact as well as the identification of portable tools to increase sensitivity are immediately relevant in this policy context.

3.2 Study design

3.2.1 Program descriptions

In our experiment, respondents see six hypothetical program descriptions. Each program description begins with a sentence outlining the broad program intent and approach, for instance a "community-based program that provides person-centered care to people with Alzheimer's Disease and Related Dementias (ADRD)". Appendix Section D shows screenshots from the full experiment, including an example program description. For 59% of respondents—generally those recruited via Evidence Leads at particular agencies—programs are catered to the agency in which the respondent works, and as such should cover topics at least broadly familiar to the respondent.[8]

3.2.2 Impact features

After the sentence introducing the program, respondents learn about the impact-relevant features of a program (henceforth called "impact features"), corresponding to each of the three following categories:

1. **Scope** - The number of people reached; that is, the number of people who have the potential to be impacted by the program.

2. **Outcome** - Whether the program impacts an intermediate outcome (for instance, click rates on an ad for a community group promoting good health habits) versus a downstream outcome (for instance, enrollment in the group, or even resulting health outcomes). The translation between intermediate and downstream outcomes is clearly presented when the outcome is intermediate. For instance, if the program outcome is click rates on an ad for a community group, the text might state that 1 in 100 people

[8]The full list of program descriptions can be found in Appendix Table G.1.

who click on the ad because of the program go on to enroll in the group who wouldn't have otherwise.

3. **Persistence** - How long the program effects last, defined as the number of years the average program recipient showed positive outcomes compared to a control group who did not receive the program, accounting for the length of the evaluation.

The three impact features are listed as separate bullets on each program description screen, as shown in Appendix Figure D.4. They are made very salient for the sake of clarity and comprehension.[9]

Importantly, each impact feature can take on a "high" or a "low" value for a particular program, and the impact features vary both within and across programs. As such, some participants are randomly assigned to see a relatively low-impact version of a program (accounting for the three impact features) while others see a higher-impact version of the same program. There are four possible combinations of impact features that respondents can see for each program, which correspond to:

1. High Scope, High Outcome, High Persistence
2. Low Scope, High Outcome, High Persistence
3. High Scope, Low Outcome, High Persistence
4. High Scope, High Outcome, Low Persistence

Each respondent is randomly assigned to see just one possible impact combination for any particular program (the exception being that two impact combinations for the same program are shown together in the Side-by-Side condition described below).

3.2.3 Program assessments

After reading about each program and its corresponding impact features, respondents assess the value of the program via a modified multiple price list approach. Specifically, as shown in Appendix Figure D.5, respondents select the maximum cost at which they think the program would be worth funding, from a semi-logarithmic list of costs ranging from "less

[9]This suggests we may overestimate baseline sensitivity to impact compared to the sensitivity we might observe in real-world policymaking contexts. Section 4.1 provides a more complete discussion of factors that may influence this estimate in either direction.

than $1,000" to $1 billion.[10][11] The semi-logarithmic scale allows us to capture a broad range of responses that correspond to what one might consider a small, medium, or large program.

This procedure identifies a respondent's indifference range, in that for any cost equal to or below the selected cost the respondent would support funding the program, while for any cost equal to or above the next option on the cost list the respondent would consider the program too expensive to fund. This assessment of program value is our primary outcome of interest.[12]

3.2.4 Estimating program impact

In order to estimate the elasticity of program assessments with respect to impact, we must generate a quantifiable measure for program impact. To do so, we multiply together the values of the three impact features (scope, outcome, and persistence), which provides an aggregate value for the person-years of impact for our downstream outcome:

$$\textbf{Impact} = \frac{\text{Number Reached} \cdot \text{Treatment Effect} \cdot \text{Years of Effect}}{\text{Probability of Intermediate Converting to Downstream Outcome}}$$

For instance, consider a program that reaches 1 million people in total (scope), increases the likelihood of achieving intermediate outcome X by 10pp (outcome), and has effects that last for one year (persistence). If 1 in 100 people who achieve intermediate outcome X because of the program ultimately achieve downstream outcome Y (which we state clearly in the program description), then we have:

[10] The precise wording of this prompt depends on the agency respondents belong to. Policymakers at the Department of Education, for instance, were asked about the maximum cost "such that you would recommend that the Department fund the program at this cost but not for any higher cost listed." Meanwhile, policymakers with the Administration for Community Living (ACL) within Health and Human Services were asked about the maximum cost to fund a *grant* because we were advised that, at ACL, employees were more likely to make decisions about which grants to award rather than what programs to directly fund.

[11] We code the "less than $1,000" response as reflecting a willingness to pay of $300, to achieve consistent log scaling across the price options. This first response option was presented as "$0" for the first 75 respondents and then changed to "less than $1,000" to avoid zero values when taking logs. Results are robust to coding this lowest-value response as $1 or $1,000 instead of $300.

[12] The stated- and revealed-preferences literatures often estimate assessments of the value of a program (for instance) via an iterative multiple price list, akin to the staircase method, in which respondents first select whether or not they think a program is worth funding for a given amount, and then answer similar questions for higher or lower amounts based on their response to the prior question (for example, Holz et al. (2021)). While this approach is well-suited to the lab, we instead use the price list presented on a single page in order to estimate a relatively narrow indifference range without burdening time-constrained policymakers with many additional pages of "yes" or "no" questions.

$$\textbf{Impact} = \frac{1{,}000{,}000 \cdot \frac{1}{10} \cdot 1}{100} = 1{,}000 \text{ person-years impacted}$$

As noted above, there are four possible combinations of our three impact features for each program. The combinations are constructed such that impact is 10 times, 100 times, and 1,000 times larger than the lowest-impact combination.[13] Aggregating our impact features into four possible combinations allows us to more easily compare assessments across different programs and ultimately identify how assessments of program value scale with program impact. Intuitively, if the measured impact of one program variation was 100 times larger than that of another and a respondent's assessments of program value scaled one-to-one with impact, then we would see that the respondent's assessment of the value of the program would also be 100 times larger.

Of note, while the aggregate measure of program impact is a useful tool for summarizing findings, all analyses presented in this paper can be (and often are) performed without aggregating across the impact features. For instance, we compare not just how assessments of program value vary with impact overall, but also how assessments vary with scope, outcome, and persistence individually.

3.2.5 Empirical framework for estimating sensitivity

Our experimental design sets the stage for an estimation of *sensitivity to impact*, or the relationship between a percent-change in program impact and a percent-change in assessments of program value. To examine the relationship between program impact and value, we first log-transform both measures. Then, to account for differences in the baseline value ascribed to different types of programs, we subtract from each assessment the average assessment reported for the lowest-impact combination for each program.

Our pre-registered regression specification uses these scaled and log-transformed measures to estimate the effect of a change in program impact on respondents' assessments of the value of the program. The regression is run within respondents i, at the level of each program p,

$$y_{ip} = \beta_0 + \beta_1 I_{ip} + \delta_i + \alpha_p + \epsilon_{ip} \tag{1}$$

where:

[13] To facilitate independent estimates of elasticity with respect to our three different impact features, the features assigned to a given impact level vary across programs; for instance, the combination that is 100 times more impactful than the lowest-impact combination for one program may be Low Scope-High Outcome-High Persistence, while another program will have a different combination corresponding to the same impact level.

- $y_{ip} \equiv$ the primary outcome of interest, i.e. the scaled and log-transformed assessment of program value, as defined in Section 3.2.3;
- $I_{ip} \equiv$ the scaled and log-transformed program impact, as defined in Section 3.2.4;
- $\delta_i \equiv$ respondent fixed effects;
- $\alpha_p \equiv$ program fixed effects.

Robust standard errors are adjusted to reflect clustering at the respondent level. β_1 measures the elasticity of assessments of program value with respect to impact.

3.2.6 Treatments

The procedure described above of using assessments of program value to estimate sensitivity to impact is fixed for every assessment in the experiment. Importantly, however, all assessment decisions are not the same. The study is broken down into three conditions, one of which serves as the Control condition, and two of which introduce decision aids aimed at simplifying the decision problem and, in turn, increasing sensitivity. The order in which these conditions appear is randomized across participants:

1. **Control:** In the Control condition, program descriptions and assessment decisions are presented with no additional information (see Appendix Figure D.5 for an example), facilitating the documentation of sensitivity to impact at baseline.

2. **Side-by-Side** In the Side-by-Side condition, two programs are presented together on one decision page rather than in isolation (see Appendix Figure D.7 for an example). In other conditions respondents see entirely distinct programs, but in the Side-by-Side condition respondents see two programs that are the same but for different impact combinations. The combinations are assigned such that one program includes the three impact features that make up the highest-impact version of this program, while the other program includes one lower-impact feature. This condition is intended to increase sensitivity to impact by facilitating comparisons across relevant impact features.

3. **Impact Calculator:** In the Impact Calculator condition the annual cost per person who achieves the downstream program outcome, but who wouldn't have without the program, is calculated for each program cost in the price list (see Appendix Figure D.8 for an example). In other words, when assessing the dollar value of a program, respondents see not only the set of possible total program costs but also the corresponding "annual cost per person impacted." This calculation is based entirely on the three impact features made available to the respondent; text on the bottom of the decision

page tells participants that this number was calculated by estimating program impact via the process described in Section 3.2.4 and then dividing each proposed program cost by this amount. This condition is intended to increase sensitivity by serving as an aid both in providing a "theory" for how to use the impact features to inform an estimate of program value and also by presenting one aggregate, more digestible metric for impact while absolving respondents of the burden of doing the math themselves.

The three conditions allow us to document the degree to which decision makers are (in)sensitive to impact-relevant features of a program (Control), as well as whether aids that aim to reduce the complexity of the decision problem (Side-by-Side and Impact Calculator) can increase sensitivity.

3.2.7 Certainty

The theory of cognitive uncertainty applied to programmatic decision-making predicts that individuals who are more certain in their decisions—perhaps due to more real-world experience, or a more developed "theory" of how to use the impact-relevant information when assessing program value—will in turn be more sensitive to impact.[14] To explore whether this holds true in our setting, we ask respondents two sets of questions that serve as a proxy for the degree of cognitive uncertainty faced when making program assessments. The first set of questions, adapted directly from Enke and Graeber (2021), appear after each control question (Appendix Figure D.6 provides an example). For these questions, respondents are first reminded of the assessment they provided for the value of a program and are then asked, "How certain are you that this is the best possible assessment, given what you have been told about the program?" These questions allow us to explore the correlation between uncertainty and sensitivity at baseline. The second set of questions come at the end of the survey, and ask respondents to assess whether they were "more, less, or equally certain" when making assessments in the "Side-by-Side" and "Impact Calculator" conditions (see Appendix Figure D.9). These questions allow us to identify whether the decision aids increase certainty in assessment decisions.

3.2.8 Experience and background

The experiment concludes with six questions assessing respondents' experience with evidence, evaluation, and programmatic decision-making. We include these questions to investigate whether policymakers with more experience with the type of decision problems in

[14] Among policymakers in our experiment, self-reported certainty is strongly predictive of more experience with evidence and evaluation, also self-reported ($p = 0.002$).

the experiment are those who find it easier to update about the information on program impact. These questions appear among a broader set of questions meant to provide insights on evidence and evaluation in government for our agency partners; as such, to prevent overburdening respondents we only present a subset of these questions in each survey.[15] Finally, we ask respondents about their office and (pay) grade in government as well as standard demographic questions eliciting their age, race, gender, and level of education.

4 Experimental Results

4.1 Documenting policymakers' sensitivity to impact

Column 1 of Table I reports the OLS estimates from Equation 1. Consistent with our first hypothesis, we see that the elasticity of policymakers' assessments of program value with respect to program impact is 0.33; that is, when program impact increases by 100%, assessments of the value of the program increase by just 33%.[16][17][18] The Control (gray) line in Figure II shows this relationship between program assessments and program impact visually. We see here that assessments of program value are distributed via a power law with a slope less than 1, in accordance with theories of cognitive misperceptions (Khaw et al., 2020).

Column 2 of Table I shows the relative sensitivity to our three impact features: persistence, scope, and outcome type. Respondents are most sensitive to persistence.[19] This may be explained by the fact that the persistence of program effects is presented in easily digestible units (days, weeks, months, or years of impact) and that low values of persistence (effects that last just a few days) may be particularly salient to respondents. Consistent with

[15] These additional agency-specific questions ask respondents to describe the impact features of a program they have been involved in implementing, the barriers they think most affect evidence utilization in practice, and recommendations to improve evidence utilization. While the responses to these mostly open-ended questions provide interesting insights for understanding the decision-making context, they were primarily included to answer questions of interest to our agency partners. See Appendix Figure D.10 for examples of these questions.

[16] Note that attenuation bias cannot explain the low elasticity estimates we observe given that our independent variables are precisely estimated and noise will only enter through the outcome, participants' assessments of program value (Frost and Thompson, 2000).

[17] Appendix Figure A.2 shows the distribution of assessments of the value of a program in the Control condition. Log assessments of program value follow a roughly normal distribution centered around $3 million, with an additional spike in assessments of "less than $1,000."

[18] Due to unforeseen technical glitches, there were a small number of instances in which a program assessment page was skipped in the experiment. As can be seen in the observation count in Table I, one respondent did not have the opportunity to make the two Control assessments as a result.

[19] This is especially interesting given that policymakers might discount the future value of a program and therefore respond relatively less to information about persistence. However, discounting is unlikely to play a substantial role given that the recent real discount rates the government recommends for use in cost-effectiveness analysis are close to zero (Office of Management and Budget, 2020).

Table I: Sensitivity to Impact at Baseline

	(1) Scaled Assessment	(2) Scaled Assessment
Scaled Impact	0.332***	
	(0.072)	
Scaled Persistence		0.583***
		(0.127)
Scaled Scope		0.238**
		(0.097)
Scaled Outcome		0.229**
		(0.093)
Observations	380	380
Median Assessment	$3 million	$3 million
Respondent FE	Yes	Yes
Program FE	Yes	Yes

This table shows the results of an OLS regression relating program impact to the two assessments of program value made by each individual in the Control condition. Column 1 reflects sensitivity to the aggregated impact of a program derived from our three impact features, as described in Section 3.2.4. Column 2 reflects sensitivity to the three independent impact features. Standard errors are in parentheses and are clustered at the respondent level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

the existing literature, we observe a low elasticity of assessments with respect to scope, or the number of people reached. Elasticity is similarly low with respect to the program's outcome. This latter object is arguably the least digestible, requiring respondents to understand and value the translation between intermediate and downstream program outcomes.

The low elasticity estimates observed in our experiment are notable for three reasons. First, the impact-relevant features of a program were made very salient in our experiment—respondents saw only one sentence describing the program in broader terms, and then saw three bullets highlighting the program's scope, outcome, and persistence. Second, many respondents were told that they were recruited specifically because of their involvement in evidence and evaluation communities in government. As such, any experimenter demand invoked by the presentation of the study should work in the direction of increasing sensitivity.[20] Finally, the respondents selected for this study can reasonably be considered experts

[20]A note on experimenter demand: It is conceivable that respondents focus more on the impact features of a program when assessing its value than they typically would outside of this experiment, leading to an overestimation of our sensitivity result at baseline. However, it is difficult to see how experimenter demand could affect our treatment effects. That is, respondents do not know which conditions are control versus treatment (or even that there is a control condition), nor is it likely that they know how to calibrate their assessments in a way that increases sensitivity in the treatment conditions, other than by intuiting more sensitive responses by relying on the decision aids, which is exactly the effect in which we are interested.

on these types of decisions, given that all respondents are in some way involved in generating, interpreting, or making decisions based on evidence about government programs at the federal level. This is therefore the group for whom we would expect the greatest degree of sensitivity. Of course, there are also reasons we might expect sensitivity to be higher in actual program funding decisions compared to the hypothetical choices in our experiment. For instance, while we capture initial judgments of program value in the experiment, policymakers often spend months or even years learning about or evaluating a program. Funding decisions are also often made in collaboration with others, and with very large stakes. Given these factors that necessarily vary across settings, the particular elasticity estimate might be higher or lower in different real-world contexts, although the treatment effects and mechanisms underpinning sensitivity should be generalizable.

Finally, one might worry that we observe this attenuated sensitivity to program impact because as impact increases, policymakers grow increasingly less certain of the program's total value.[21] For instance, perhaps policymakers are more skeptical of the validity of a program that affects millions of people given documented threats to the scalability of experimental results (Al-Ubaydli et al., 2017). While we try to mitigate this concern by emphasizing in the instructions (in bold font) that respondents "should assume that the program would be implemented exactly as described in this survey," we cannot altogether account for this possibility in the experimental design. However, the data suggest that this is unlikely to explain our results. For one, the variance in respondents' log assessments of program value is similar across changes in program impact.[22] That is, responses do not grow noisier with changes in program impact. We can also use our measure of certainty (discussed further in Section 4.3.2) to look directly at whether certainty decreases when program impact increases. As shown in Appendix Figure A.3, we see no evidence of this.

4.2 Treatment effects on sensitivity

With our Control condition alone, we cannot determine whether the complexity of the decision problems is playing a role in limiting sensitivity to impact. Because simplifying these decision problems should only increase sensitivity to impact for those affected by bounded rationality constraints, a test of the efficacy of our two decision aids helps to shed light on this question. We estimate the impact of the decision aids on sensitivity by interacting

[21] Indeed, DellaVigna et al. (2022) find that policy practitioners are relatively reserved in their predictions of program treatment effects compared to academics.

[22] At baseline, $\sigma = 3.42$ for the lowest-impact programs, $\sigma = 3.77$ for programs 10 times more impactful than the lowest-impact programs, $\sigma = 3.44$ for programs 100 times more impactful than the lowest-impact programs, and $\sigma = 3.37$ for programs 1,000 times more impactful than the lowest-impact programs. The variance in log assessments is also similar across treatment conditions.

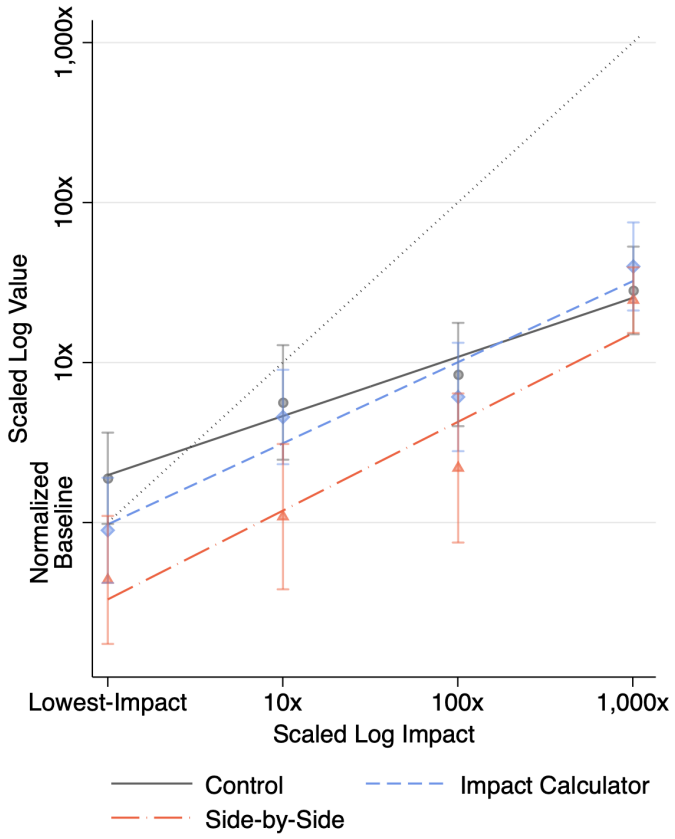


FIGURE II: Sensitivity Across Study Conditions

This figure shows the relationship between program impact and policymakers' assessments of program value in the control and two treatment conditions. The x-axis indicates the program impact, for each of the four possible impact combinations. The y-axis indicates the assessment of program value, compared to the average assessment provided for the lowest-impact variant of a program (computed separately for each program). Each point reflects the average program assessment for the corresponding impact level in a given condition, alongside 95% confidence intervals. Both impact and assessments are scaled according to the procedure described in Section 3.2.5.

an indicator for our treatment conditions with program impact. In support of our second hypothesis, both decision aids have a large and statistically significant impact on sensitivity to impact. As shown in the regression estimates in Table II and also visually in Figure II, the Side-by-Side presentation of program information increases our elasticity estimate by 79% (0.26 on a base of 0.33) and the Impact Calculator increases elasticity by 60% (0.20 on a base of 0.33). We can clearly reject a null effect of both treatments together as well as each treatment independently (joint: $p = 0.004$, Side-by-Side: $p = 0.001$, Impact Calculator: $p = 0.024$). As shown by the corrected p-values in square brackets in Column 2 of Table II, these treatment effects are robust to multiple hypothesis corrections. Of note, there is no

Table II: Impact of Treatments on Sensitivity

	(1) Scaled Assessment	(2) Scaled Assessment
Pooled Treatment X Scaled Impact	0.198***	
	(0.069)	
Pooled Treatment	-1.223***	
	(0.314)	
Side-by-Side X Scaled Impact		0.260***
		(0.077)
		[0.002]
Impact Calculator X Scaled Impact		0.196**
		(0.086)
		[0.044]
Side-by-Side		-1.895***
		(0.412)
Impact Calculator		-0.840**
		(0.362)
Baseline Sensitivity	0.33	0.33
Respondent FE	Yes	Yes
Program FE	Yes	Yes
Observations	1130	1130

This table shows the results of Equation 1 with additional (interacted) indicators for decisions made in treatment conditions to estimate the causal impact of the two treatments on sensitivity. Column 1 estimates the effect of being in any treatment condition, while Column 2 estimates the independent impact of each treatment. FWER-adjusted p-values are in square brackets in Column 2; the process for applying these corrections is described in more detail in Appendix Section C. Standard errors are in parentheses and are clustered at the respondent level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

statistically significant difference in the effects of the Side-by-Side presentation compared to the Impact Calculator ($p = 0.580$).

Interestingly, both decision aids—and particularly the Side-by-Side condition—operate by decreasing assessments for lower-impact programs. The negative coefficients on the two treatment indicators in Table II point to this effect, while Appendix Figure A.4 further elucidates the underpinning mechanics when looking at the raw value assessments. In particular, we see that the difference between treatment and control is most pronounced for the lowest-impact programs. As program impact increases, control and treatment assessments converge, until they are nearly indistinguishable for the highest-impact programs. In other words, the decision aids appear to operate by helping respondents to recognize when a program is relatively low-impact.

Observing when the Side-by-Side condition is and is not effective provides further evidence regarding the operalization of the treatments. Recall that in the Side-by-Side condition the highest-impact program (in which all three impact features take on their "high" values) was always presented next to one of its three lower-impact counterparts (in which one of the three impact features takes on a "low" value). As can be seen in Figure II, we find that in the Side-by-Side condition respondents are very attuned to which of the two programs on the screen is higher-impact ($p < 0.001$), but they differentiate relatively less between the three lower-impact levels. In other words, the Side-by-Side presentation appears to nudge respondents to attend to "high" versus "low," providing support for ordinal comparisons, but to neglect to some extent the more subtle differences across programs.

Finally, the experimental design allows us to explore whether sensitivity increases when participants make an assessment in the Control condition after exposure to a decision aid. As shown in Appendix Table A.2, we do not observe such learning effects: The (randomized) order in which the decision aids are presented does not impact sensitivity in the Control condition.[23] This suggests that while these interventions can affect responses to the particular decision to which they are applied, they do not translate to increases in sensitivity more generally in subsequent assessments.

4.3 Predictors of sensitivity

4.3.1 Individual-level differences

In order to explore the degree to which individuals differ in terms of their estimated sensitivity, we look across the six assessments made by each respondent to construct an individual-level measure of sensitivity. Figure III plots the sensitivity observed among each quartile. This figure shows that the top quartile in terms of sensitivity updates their assessments of program value roughly one-to-one in response to changes in impact (the green line is close to a 45-degree line). Meanwhile, the bottom quartile is almost perfectly *insensitive* to impact—assessments of the value of the lowest-impact programs look almost identical to assessments of programs that are 1,000 times more impactful. We turn to our correlational evidence to look at the factors that predict these individual-level differences, and whether, as per our third hypothesis, they are consistent with mechanisms related to bounded rationality.

[23] In the general public sample we also observe even more robust evidence, given the larger sample size, for this lack of learning effects.

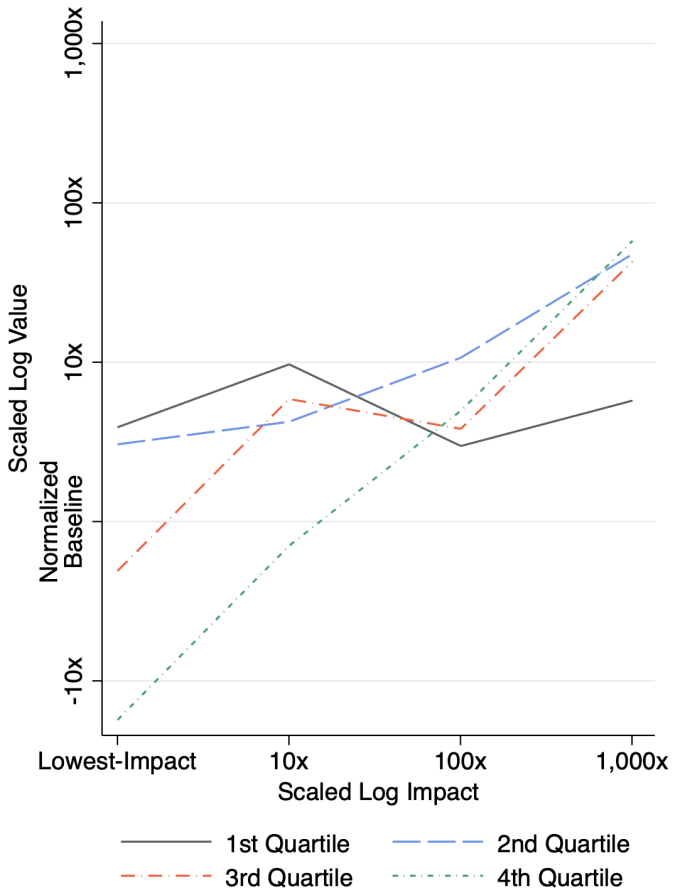


FIGURE III: Individual-Level Sensitivity Quartiles

This figure plots the relationship between program assessments and program impact for respondents by sensitivity quartile, estimated across the six assessments made by each respondent. Both impact and assessments are scaled according to the procedure described in Section 3.2.5.

4.3.2 Certainty

We first look at our measure of certainty, in which respondents indicate on a scale from 0 to 100 how certain they are that they gave the "best possible assessment" in each Control decision, given what they were told about the program. Consistent with the hypothesis that those who self-report more certainty may respond more to impact because they have a clearer idea of how to incorporate the information about the three impact features, we see in Column 5 of Table III a positive relationship between sensitivity and our measure of certainty ($p = 0.015$). Intuitively, this suggests that respondents are aware of the difficulty of mapping the information they are receiving onto program assessments.

We also see some evidence that the two treatments increase certainty. While the certainty scales are only presented in the Control condition to reduce the overall length of the

experiment, respondents were asked after completing all program assessments whether they felt more certain about their decisions when presented with each decision aid. 41% reported feeling more certain in the Side-by-Side condition while only 14% reported feeling less certain. For the Impact Calculator condition, 46% of respondents reported feeling more certain on decision screens with an Impact Calculator while only 13% reported feeling less certain.

Table III: Heterogeneities in Sensitivity

	Experience				Other Factors		
	(1)	(2)	(3)	(4)	(5)	(6)	
	Eval Exp	Grade	Familiar	**Index**	Confidence	Resp Time	**All**
Factor X Scaled Impact	0.036	0.050	0.185**	0.268***	0.006**	0.000	
	(0.023)	(0.035)	(0.073)	(0.072)	(0.003)	(0.001)	
Evidence Exp (0-10)	-0.085						0.046**
	(0.121)						(0.023)
Grade (GS 11-15)		-0.248					0.060
		(0.216)					(0.038)
Familiar Program Domain			-0.167				0.198*
			(0.474)				(0.120)
Confidence (1-100)					-0.032		
					(0.011)		
Response Time (Quartiles)						-0.004	0.001
						(0.008)	(0.002)
Observations	719	873	1130	582	376	1130	582
Program FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Scaled Impact	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Treatment Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

This table shows the relationship between participant characteristics and sensitivity. The first row shows the coefficients of interest, i.e. the interaction between these characteristics and the scaled program impact. Experience with evidence and evaluation is the average of six questions, measured on a scale from 0 to 10. Grade in government, which we use as a proxy for relevant work experience, is self-reported by participants. Missing observations for the two measures of experience are due to the fact that not all respondents completed the follow-up survey to report own characteristics and also to a feature of the experiment that randomly assigned some follow-up questions to reduce total survey length. Familiarity with program domains is an indicator equal to one when the participant is assigned to a survey that only includes programs relevant to their policy area of expertise. Index is an average of the three standardized vectors of experience. Certainty reflects self-reported certainty in program assessments, measured on a scale from 0 to 100. Response time is the average response time on all program assessments, broken down by quartile indicators. The final column reports these interactions together in one regression, excluding certainty which is only captured for Control assessments. As in Table II, controls for program impact and treatment conditions are also included. Standard errors are in parentheses and are clustered at the respondent level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.3.3 Other predictors

Columns 1-4 of Table III provide additional evidence relevant to the individual-level characteristics that relate to greater sensitivity. Columns 1 and 2 show a directionally positive but statistically insignificant relationship between sensitivity and self-reported experience with evidence and evaluation as well as grade in government.[24] In Column 3, we see that individuals who are more familiar with the program domains they encounter in the survey—i.e. those who received a survey specifically catered to the policy area in which they work—are statistically significantly more sensitive to program impact ($p = 0.013$). If we create an index (Column 4) averaging the three standardized characteristics that proxy for relevant real-world experience, we see a robust positive relationship between this index and sensitivity ($p < 0.001$), indicating that experience likely plays a role in predicting sensitivity.

Finally, Column 6 indicates that there is no statistically significant relationship between sensitivity and response time across the six program assessments ($p = 0.706$). If we think of response time as a proxy for attention, this suggests that sensitivity to impact is not merely an artifact of more or less attentive participants in the experiment—more subtly, an *understanding* of what to do with the impact-relevant information appears to be key.

In all, the correlational data provide additional evidence pointing to the role that the difficulty of mapping complex impact-relevant information onto total dollar value assessments plays in determining overall sensitivity to impact.

5 General Public Experiment

To shed light on the generalizability of our findings in the policymakers sample and inform the mechanisms underlying sensitivity in a larger sample, we replicated our experiment among a representative sample of the U.S. public.

5.1 Design

We recruited a representative sample of 500 U.S. citizens, based on 2019 Census data (U.S. Census Bureau, 2019), via the online platform Prolific. Participants were paid $2.50 for completing the survey.[25] The survey design is similar to the version developed for policymakers: After seeing the same overall set of program descriptions, participants make assessments

[24] If we only include a question on experience with "interpreting evidence/program evaluations" rather than the pre-registered six-item index that also includes arguably less relevant aspects of experience, then the relationship between self-reported experience and sensitivity is statistically significant ($p = 0.036$).

[25] Because the response time was longer than that observed in pilot data, we also gave participants a $0.50 bonus after all participants had completed the survey to compensate them for their additional time.

of program value across two Control conditions, two Impact Calculator conditions, and two Side-by-Side conditions. Participants are also exposed to an additional "Joint" condition, which combines the two treatments (see Appendix Figure F.2).

Because survey length is less of a constraint than in the policy setting, we include some additional questions. First, participants are asked to self-report certainty in their responses after both control and treatment assessments. Second, participants work through a four-question module to assess their numeracy after completing their assessments.[26] Third, we include more sensitive questions, notably about participant politics and income, which would not have been appropriate to ask in the government context.

5.2 Experimental results

500 participants completed the survey, and 94% of these passed a simple comprehension check. The results presented here include all 500 participants; results are robust to exclusion.

5.2.1 Sensitivity to impact

Table IV: Sensitivity to Impact at Baseline - General Public Sample

	(1) Scaled Assessment	(2) Scaled Assessment
Scaled Impact	0.210***	
	(0.043)	
Scaled Persistence		0.264***
		(0.064)
Scaled Scope		0.317***
		(0.054)
Scaled Outcome		0.084*
		(0.051)
Observations	1000	1000
Median Assessment	$1 million	$1 million
Respondent FE	Yes	Yes
Program FE	Yes	Yes

This table shows the results of an OLS regression relating program impact to the two assessments of program value made by each individual in the Control condition. Column 1 reflects sensitivity to the aggregated impact of a program derived from our three impact features, as described in Section 3.2.4. Column 2 reflects sensitivity to the three independent impact features. Standard errors are in parentheses and are clustered at the respondent level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

[26] These questions were adapted from a longer numeracy assessment included in Kahan et al. (2012).

Consistent with their relative lack of experience with these types of decision problems, we observe that the general public is relatively less sensitive to program impact compared to policymakers in the U.S. government, although this difference is not statistically significant ($p = 0.147$). As can be seen in Table IV, the elasticity estimate among the general public is just 0.21 while among policymakers it is 0.33.

5.2.2 Treatment effects on sensitivity

Table V: Impact of Treatments on Sensitivity - General Public Sample

	(1) Scaled Assessment	(2) Scaled Assessment	(3) Scaled Assessment
Pooled Treatment X Scaled Impact	0.208***	0.262***	
	(0.042)	(0.040)	
Pooled Treatment	-1.090***	-1.524***	
	(0.192)	(0.182)	
Side-by-Side X Scaled Impact			0.254***
			(0.047)
Impact Calculator X Scaled Impact			0.172***
			(0.049)
Joint Treatment X Scaled Impact			0.394***
			(0.047)
Side-by-Side			-1.387***
			(0.243)
Impact Calculator			-0.877***
			(0.218)
Joint Treatment			-2.566***
			(0.243)
Baseline Sensitivity	0.21	0.21	0.21
Respondent FE	Yes	Yes	Yes
Program FE	Yes	Yes	Yes
Joint Treat Included	No	Yes	Yes
Observations	3000	4000	4000

This table shows the results of Equation 1 with additional (interacted) indicators for decisions made in treatment conditions to estimate the causal impact of the two treatments on sensitivity. Column 1 estimates the effect of being in either the Side-by-Side or Impact Calculator condition and excludes the Joint Treatment for the purposes of comparison with the policymakers experiment. Column 2 estimates the effect of being in any condition, including the Joint Treatment. Column 3 estimates the independent impact of each treatment. Standard errors are in parentheses and are clustered at the respondent level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

As shown in Table V, we also see large treatment effects among the general public—while the two treatments increase sensitivity by an average of 60% in the policymakers sample, they increase sensitivity by close to 100% in the representative sample of U.S. citizens ($p < 0.001$).

This is in part mechanical; because baseline sensitivity is lower in this population, there is a higher ceiling under which treatment effects may operate. Still, larger treatment effects in this sample were by no means inevitable. The opposite result, in which participants were so poorly attuned to evidence-based information that they did not understand or update based on the new methods of presenting information, would have also been reasonable to observe.

Appendix Figure B.1 indicates that we see the same pattern of divergence in assessments between treatment and control at the lower impact levels as well. That is, when participants see the highest-impact version of a program, assessments are similar across treatment and Control conditions. However, for lower-impact programs, the treatment conditions tend to elicit systematically lower assessments of program value. We also again see that the Side-by-Side condition operates primarily by highlighting the difference between the highest-impact program and all three lower-impact combinations for a program.

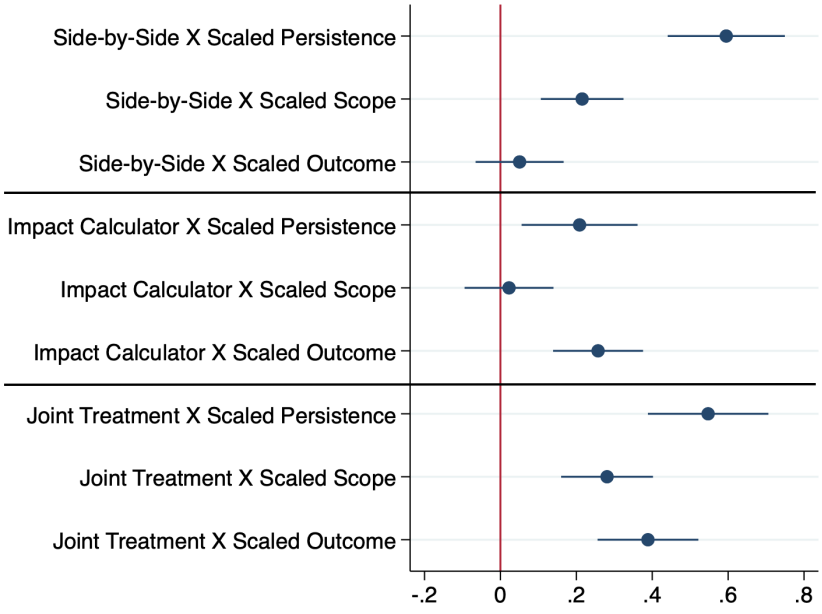


FIGURE IV: Treatment Effects by Impact Feature - General Public Sample

This figure plots the coefficients from a modified version of Equation 1, which interacts the three treatments included in the general public experiment with the log-scaled impact level for each impact feature. Error bars reflect 95% confidence intervals.

Rather remarkably, we see in Columns 2 and 3 of Table V that the "Joint Treatment," which combines the features of the Side-by-Side and Impact Calculator conditions on one decision page, increases sensitivity by 188% ($p < 0.001$). That is, it appears that the two treatments have roughly additive effects. Figure IV helps to explain this by illustrating the impact of the treatments on sensitivity to each of our three impact features: persis-

tence, scope, and outcome type.[27] While the effects are directionally positive across the board, we see that the Side-by-Side presentation only has a statistically-significant impact on sensitivity to program persistence and scope, while the Impact Calculator only clearly increases sensitivity to outcome type and persistence. This is consistent with how the two treatments intuitively operate: The Side-by-Side condition highlights differences between easily-understood metrics, which both the number of people reached and the program duration plausibly represent. The Impact Calculator, meanwhile, helps to clarify differences in impact that may be relatively more difficult to digest without the translation into an easily-digestible metric, which, intuitively, may particularly pertain to the outcome type. Given the somewhat complementary effects of the two treatments, when combined via the Joint Treatment, we see consistently strong treatment effects across all impact features.

5.2.3 Predictors of sensitivity

Mimicking the parallel exercise in the policymakers sample, Appendix Figure B.2 plots sensitivity to impact at the individual level. We see here that even the top quartile in terms of sensitivity in the general public does not update their assessments one-to-one in response to changes in impact.

In contrast to our policymakers sample, Column 1 of Appendix Table B.1 shows that among the general public these differences in sensitivity are not explained by differences in self-reported certainty. It may be the case that policymakers are more aware of their "cognitive uncertainty." Indeed, self-reported certainty for control assessments is 19% higher among the general public compared to policymakers. Or, perhaps the general public's certainty assessments are driven more by the *type* of program—for instance, an education program that supports students with disabilities—rather than its impact per se. This explanation is consistent with the fact that we observe no impact of the treatments on the assessment-specific certainty questions, despite the fact that participants do report that the decision aids increased their certainty when asked specifically about this at the end of the survey.

On the other hand, Appendix Table B.1 does show differences in sensitivity by respondent-level characteristics. Notably, the percent of questions answered correctly on our four-question numeracy assessment is strongly predictive of sensitivity to impact ($p < 0.001$). This effect holds up when controlling for other relevant predictors. This is consistent with our third hypothesis that bounded rationality is an important underlying mechanism in that more numerate individuals are plausibly better able to mentally map evidence onto a program value assessment. Although this analysis is exploratory, we also see some evidence of

[27] Appendix Figure A.5 shows the comparable figure for the policymakers sample; the data are consistent with the effects observed among the general public but are, naturally, less well-powered.

differences in sensitivity by political ideology. Individuals who lean conservative—according to a 0-7 scale where 0 indicates "extremely liberal" and 7 indicates "extremely conservative"—are less sensitive to changes in impact ($p = 0.021$). There is also suggestive (but statistically insignificant) evidence that conservatives provide lower average assessments for the value of a government program.

5.2.4 Incentivized predictions

Because our main assessment decisions elicit individuals' beliefs and values and therefore do not have verifiable answers, we cannot use incentives to increase our confidence that the stated answers accurately reflect respondents' actual beliefs. However, we run an additional variant of our general public experiment in which respondents are asked to predict others' beliefs, such that questions do have a clear correct answer. In this survey 250 new respondents see the same selection of program descriptions, and they are asked to predict the "most typical answer provided by other survey respondents." A bonus payment of 20 cents is paid for each of eight correct predictions. As can be seen in Appendix Table B.2, both the point estimate for predicted sensitivity at baseline as well as the predicted treatment effects are similar to those observed among the general public in the main experiment.[28] This indicates that the low elasticity of assessments with respect to program impact that we observe is unlikely to simply be an artefact of under-attention due to a lack of incentives. Note, however, that this additional evidence is merely suggestive given that sophisticated participants could conceivably anticipate such under-attention when making predictions.

We also ask participants in this experiment to indicate the degree of sensitivity they expect among the general public and policymakers, as well as how sensitive they think responses "should" be. In particular, at the end of the experiment we ask, "By what factor do you think people's assessments of the value of a program (should) change when its impact increases by a factor of 10?" Appendix Figure B.3 presents the results, from which several insights emerge. First, we see that both the median and modal response to the question about what the scaling factor *ought* to be is 10; that is, a typical participant thinks scaling assessments one-to-one with respect to changes in program impact is the optimal response. This provides further evidence that respondents would provide more sensitive responses if they were able, and that the observed low elasticity at baseline is unlikely to simply reflect

[28]Note that we also included a similar question at the end of the survey for policymakers, but in this case only for the Control condition. While we were not permitted to offer financial incentives for correct answers in the policymaking setting, as a more subtle social incentive respondents were told that we would follow up with an email in which their response to this question would be compared to the typical response in the experiment. Rather than documenting an increase in sensitivity in response to incentives, which we might imagine if sensitivity were lower due simply to under-attention, we saw somewhat lower sensitivity in response to this question compared to the main assessments.

preferences alone. Second, participants similarly over-estimate sensitivity across the general public and policymakers samples: The median prediction for both populations is that when impact scales up by a factor of 10, assessments increase by a factor of 9; that is, participants predict an elasticity estimate of 0.9. Less than 10% of participants predict an elasticity estimate that is equal to or less than that observed in our experiments for either population, even after having seen a version of the experiment themselves.

6 Discussion and conclusion

Every year, policymakers in the U.S. government are entrusted with allocating close to $7 trillion (Department of the Treasury, 2020). Increasingly, the expectation is that policymakers will make these decisions based on evidence to ensure that federal tax dollars are allocated to the highest-impact programs. Therefore, the process by which these allocation decisions are made—and in particular whether they are affected by the complexity of the decision problems—has the potential to substantially affect the resources and opportunities available to the broader public. Our findings contribute to an understanding of this decision-making process by providing insights into how policymakers respond to evidence about program impact when assessing the value of government programs.

We use a lab-in-the-field experiment among high-ranking U.S. policymakers to document a limited sensitivity to impact. When program impact increases by 100%, the value individuals ascribe to a program increases by 33%. In a complementary experiment among a representative sample of U.S. citizens, we find that the elasticity of assessments of program value with respect to impact is 0.21, compared to 0.33. Policymakers are more sensitive to evidence about impact when they have more experience and are more certain in their assessments of program value, and in the general public numeracy is correlated with sensitivity.

Our experiment also identifies decision aids—in particular, the presentation of two similar programs Side-by-Side, along with an Impact Calculator that translates the total cost of a program into an annual cost per person impacted—that substantially increase sensitivity to impact among policymakers as well as the general public. Both decision aids are designed to simplify the translation of information about program impact into assessments of program value. As such, the large effects of these interventions point to the role of bounded rationality in limiting sensitivity to impact.

One caution to relying too heavily on decision aids in practice hinges on the importance of the quality of the inputs. In this study, for instance, we see suggestive evidence in the policymakers sample that the decision aids play a larger role in increasing sensitivity to persistence when policymakers are assessing a program for which the persistence of effects is

relatively less important ($p = 0.109$).[29] Such effects indicate that *evaluability bias* may play a role in this setting: When the framing of information makes a particular component easier to evaluate, people will put more weight on that component (Exley, 2020; Hsee, 1996). As such, tools to increase sensitivity need not always be welfare-maximizing, and attention to the selected inputs is warranted.

Practically, our intent is for the insights developed in this paper to serve as aids to researchers and evaluators looking to effectively disseminate the results of program evaluations. Given that we do not observe learning effects after exposure to the interventions in our experiment—sensitivity does not increase in the Control condition when participants have already been exposed to the treatments—we recommend incorporating decision aids directly in dissemination materials rather than using these tools to try to train decision-making. The large effects of the Side-by-Side condition suggest that the timing of program funding decisions likely also matters; assessments about several programs made together on an appointed day, for instance, are likely to be more calibrated to impact than assessments made independently for each program in isolation. Finally, our experimental method of estimating sensitivity could also be incorporated into work exploring sensitivity in any number of applied settings.

This paper points to an area of research applying insights from behavioral economics to policy-relevant decision-making that is still relatively under-explored. Future work could expand on the mechanisms and barriers underlying how people respond to impact-relevant information, how these responses vary in different contexts, and additional tools to improve evidence utilization. With these considerations in mind, we hope this paper serves to aid our understanding of environments where more effective decision-making can ultimately lead to more lives saved, less wasteful spending, and improved well-being for those affected by the decisions.

[29] We compare relatively short-run effects across programs for which the main benefits are achieved in the long-run to programs for which improved outcomes have real-time implications for participants. For instance, an education program that aims to improve test scores represents a program for which the shorter-run persistence of effects is relatively less important; if a treatment student's math scores were no different than a control student's scores 5 years after an intervention, it matters less whether the scores were improved for a semester versus a year.

References

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

Alós-Ferrer, C., E. Fehr, and N. Netzer, "Time will tell: Recovering preferences when choices are noisy," *Journal of Political Economy*, 2021, *129*, 1828–1877.

Banuri, S., S. Dercon, and V. Gauri, "Biased policy professionals," *The World Bank Economic Review*, 2019, *33*, 310–327.

Benjamin, D., "Errors in probabilistic reasoning and judgmental biases," in D. Bernheim, S. DellaVigna, and D. Laibson, eds., *Elsevier Press*, Routledge, 2019.

Bergman, P., J. Lasky-Fink, and T. Rogers, "Simplification and defaults affect adoption and impact of technology, but decision makers do not realize it," *Organizational Behavior and Human Decision Processes*, 2020, *158*, 66–79.

Bohnet, I., A. van Geen, and M. Bazerman, "When performance trumps gender bias: Joint vs. separate evaluation," *Management Science*, 2016, *62*, 1225–1234.

Boyce-Jacino, C., E. Peters, A. Galvani, and G. Chapman, "Large numbers cause magnitude neglect: The case of government expenditures," *Proceedings of the National Academy of Sciences*, 2021, *119.*

Christensen, J. and D. Moynihan, "Motivated reasoning and policy information: politicians are more resistant to debiasing interventions than the general public," *Behavioral Public Policy*, 2020, pp. 1–22.

Crowley, D., J. Scott, E. Long, L. Green, A. Israel, L. Supplee, E. Jordan, K. Oliver, S. Guillot-Wright, B. Gay, R. Storace, N. Torres-Mackie, Y. Murphy, S. Donnay, J. Reardanz, R. Smith, K. McGuire, E. Baker, A. Antonopoulos, M. McCauley, and C. Giray, "Lawmakers' use of scientific evidence can be improved," *Proceedings of the National Academy of Sciences*, 2021, *118.*

DellaVigna, S., W. Kim, and E. Linos, "Bottlenecks for evidence adoption," *NBER Working Paper 30144*, 2022.

Department of the Treasury, "Data Lab," 2020.

Dickert, S., D. Vastfjall, J. Kleber, and P. Slovic, "Scope insensitivity: The limits of intuitive valuation of human lives in public policy," *Journal of Applied Research in Memory and Cognition*, 2015, *4*, 248–255.

Enke, B. and T. Graeber, "Cognitive uncertainty," *Working Paper*, 2021.

Evangelidis, I. and B. Van den Bergh, "The number of fatalities drives disaster aid: Increasing sensitivity to people in need," *Psychological Science*, 2013, *24*, 2226–2234.

Executive Office of the President, "Restoring trust in government through scientific integrity and evidence-based policymaking," *Federal Register*, 2021, *86*.

Exley, C., "Using charity performance metrics as an excuse not to give," *Management Science*, 2020, *66*, 553–563.

Frost, C. and S. Thompson, "Correcting for regression dilution bias: Comparison of methods for a single predictor variable," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 2000, *163*, 173–189.

Gabaix, X., "Behavioral inattention," in D. Bernheim, S. DellaVigna, and D. Laibson, eds., *Elsevier Press*, Routledge, 2019.

Haskins, R. and G. Margolis, *Show me the evidence: Obama's fight for rigor and results in social policy*, Brookings Institution Press, 2014.

Hill, H. and D. Briggs, "Education leaders' knowledge of causal research design: A measurement challenge," *Annenberg Institute at Brown University EdWorkingPaper*, 2020.

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

Holz, J., R. Jimenez-Duran, and E. Laguna-Muggenburg, "Estimating repugnance toward price gouging with incentivized consumer reports," *SSRN Working Paper*, 2021.

H.R.4174, "Foundations for evidence-based policymaking act of 2018," 2019, *115th Congress.*

Hsee, C., "The evaluability hypothesis: An explanation of preference reversals between joint and separate evaluations of alternatives," *Organizational Behavior and Human Decision Processes*, 1996, *67*, 247–257.

_ **and Y. Rottenstreich**, "Music, pandas, and muggers: On the affective psychology of value," *Journal of Experimental Psychology: General*, 2004, *133*, 23–30.

_ **, G. Loewenstein, S. Blount, and M. Bazerman**, "Preference reversals between joint and separate evaluations of options: A review and theoretical analysis," *Psychological Bulletin*, 1999, *125*, 576–590.

Kahan, D., E. Peters, M. Wittlin, P. Slovic, L.L. Ouellette, D. Braman, and G. Mandel, "The polarizing impact of science literacy and numeracy on perceived climate change risks," *Nature Climate Change*, 2012, *2*, 732–735.

Kahneman, D. and Knetsch, "Valuing public goods: The purchase of moral satisfaction," *Journal of Environmental Economics and Management*, 1992, *22*, 57–70.

Khaw, M.W., Z. Li, and M. Woodford, "Cognitive imprecision and small-stakes risk aversion," *The Review of Economic Studies*, 2020, pp. 1–35.

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

Lugo-Gil, J., D. Jean-Baptiste, and L. Jaramillo, "Use of evidence to drive decision-making in government," *Mathematica Policy Research*, 2019.

Mayar, S., R. Shah, and A. Kalil, "How cognitive biases can undermine program scale-up decisions," in J. List, D. Suskind, and L. H Supplee, eds., *The Scale-up Effect in Early Childhood and Public Policy: Why interventions lose impact at scale and what we can do about it*, Routledge, 2021.

Mehmood, S., S. Naseer, and D. Chen, "Training policymakers in econometrics," *Working Paper*, 2021.

Moynihan, D. and S. Lavertu, "Does involvement in performance management routines encourage performance information use? Evaluating GPRA and PART," *Public Administration Review*, 2012, *72*, 592–602.

_ **and S. Pandey**, "The big question for performance management: Why do managers use performance information?," *The Journal of Public Administration Research and Theory*, 2010, *20*, 849–866.

Nakajima, N., "Evidence-based decisions and education policymakers," *Working Paper*, 2021.

Natow, R., "Research utilization in higher education rulemaking: A multi-case study of research prevalence, sources, and barriers," *Education Policy Analysis Archives*, 2020, *29*, 1–36.

Office of Management and Budget, "Discount rates for cost-effectiveness, lease purchase, and related analyses," *0MB Circular No. A-94*, 2020.

Saiewitz, A. and M. Piercey, "Too big to comprehend? A research note on how large number disclosure format affects voter support for government spending bills," *Behavioral Research in Accounting*, 2019, *32.*

Simon, H., "A behavioral model of rational choice," *Quarterly Journal of Economics*, 1955, *69*, 99–118.

Small, D., G. Loewenstein, and P. Slovic, "Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims," *Organizational Behavior and Human Decision Processes*, 2007, *102*, 143–153.

Tversky, A. and D. Kahneman, "Judgment under uncertainty: Heuristics and Biases," *Science*, 1974, *185*, 1124–1131.

U.S. Census Bureau, "Selected population profile in the United States," *American Community Survey*, 2019.

Vivalt, E., A. Coville, and KC Sampada, "Weighing the evidence: Which studies count?," *Working Paper*, 2021.

_ and _ , "How do policy-makers update their beliefs?," *Working Paper*, 2021.

Vought, R., "Evidence-based policymaking: Learning agendas and annual evaluation plans," *Office of Management and Budget, Executive Office of the President*, 2021.

World Bank Group, "Mind, society, and behavior," *World Development Report*, 2015.

A Policymakers Sample - Additional Results

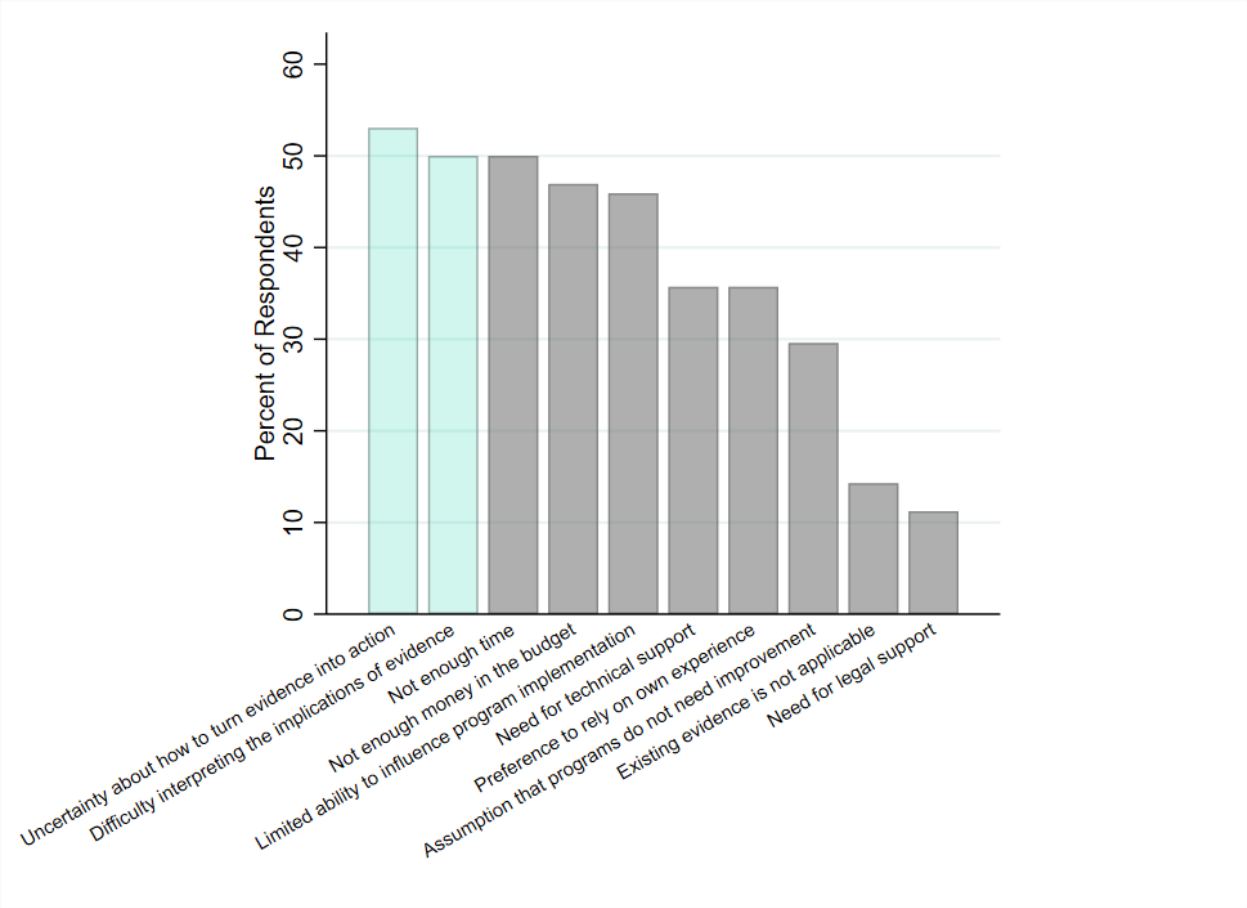


FIGURE A.1: Barriers to Evidence Utilization

At the end of the main experiment, policymakers were asked to indicate "which of the following barriers do you think interfere with the process of using evidence for programmatic decision making at your agency?" This figure shows the percent of respondents who selected each of ten barriers. Policymakers could select any number of the ten barriers, which were presented in random order. The full text policymakers saw for each barrier is as follows: "Not enough money in the budget"; "Not enough time"; "Limited ability to influence program implementation"; "Uncertainty about how to turn evidence into action"; "Difficulty interpreting the implications of evidence-based recommendations"; "Preference to rely on own experience rather than evaluations"; "Assumption that the existing programs do not need improvement"; "Need for technical support"; "Need for legal support"; "Existing evidence is not applicable to the program under consideration." Respondents could also select "other" and specify an additional barrier; 20 respondents did so. The end-of-survey questions were randomly assigned to respondents to avoid overburdening all participants with the full list of questions, such that 98 policymakers answered this particular question.

Table A.1: Policymaker Characteristics

Agency	Evidence Exp 0-10	Avg Grade	N
Department of Education	5.8	GS-14	54
Health and Human Services	6.1	GS-14	44
General Services Administration	4.9	GS-14	14
USAID	6.4	NA	9
Department of Justice	5.0	GS-14	9
Other	5.8	GS-14	61
Total	5.7	GS-14	191

This table presents summary statistics for policymakers, by the agency to which they belong. Experience with evidence and evaluation is the average of six survey questions, measured on a scale from 0 to 10. Grade in government reflects the policymakers' General Schedule Grade, which is capped at GS-15, and is self-reported by participants. The sample is composed of 191 participants in total.

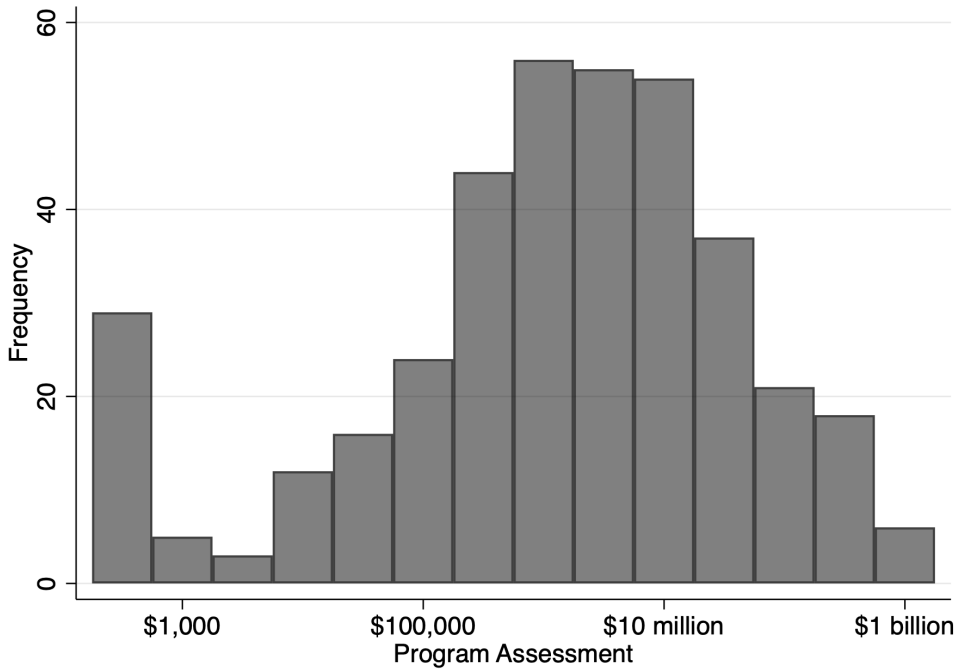


FIGURE A.2: Distribution of Control Assessments

This figure plots the distribution of assessments of the raw dollar value of programs provided in the Control condition.

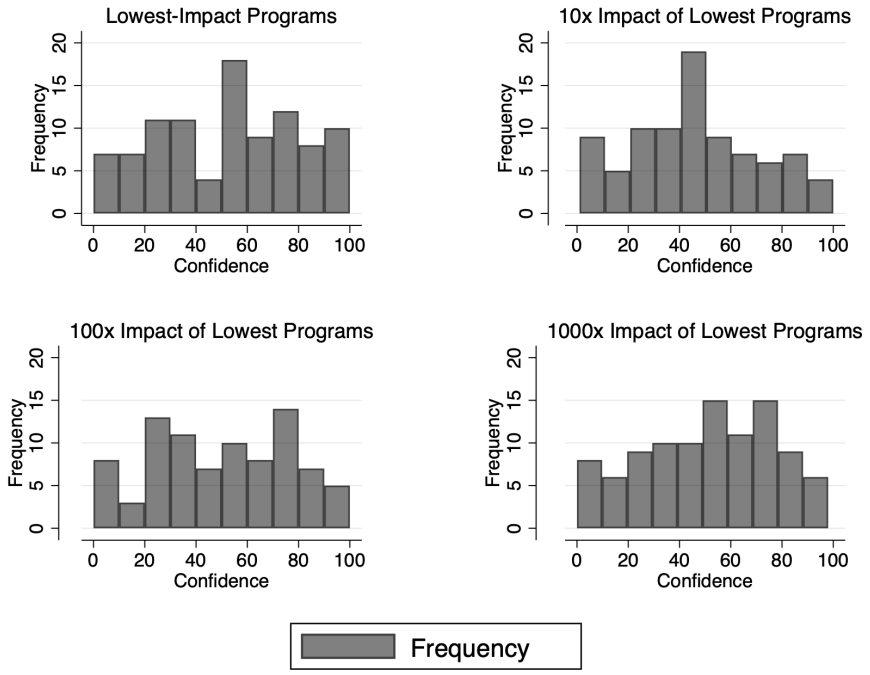


FIGURE A.3: Distribution of Certainty Elicitations

This figure plots the distributions of responses to the survey question asking respondents to indicate on a scale from 0 to 100 how certain they are that they gave the "best possible assessment" in each control decision, given what they were told about the program. The distributions are presented separately for each program impact level.

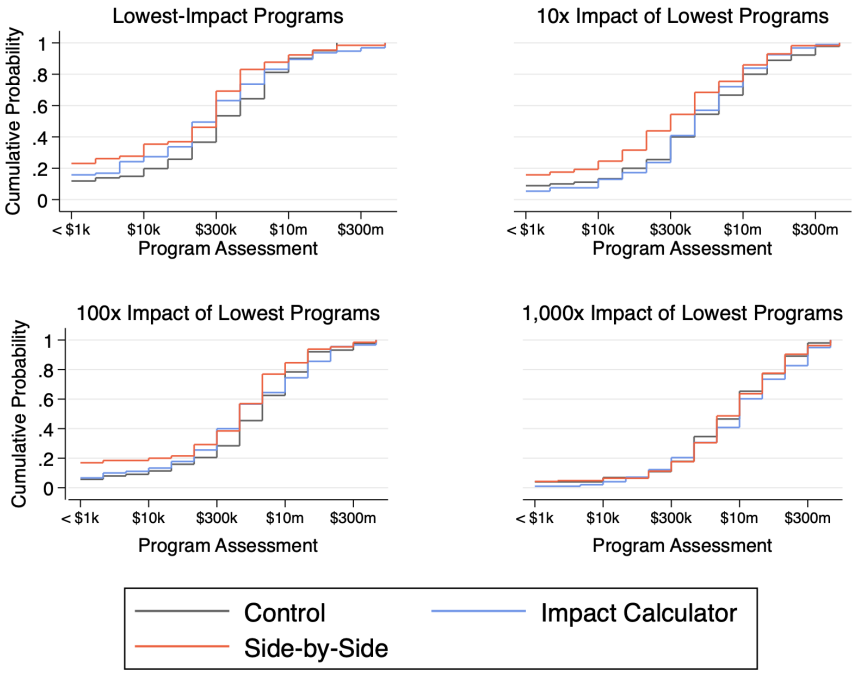


FIGURE A.4: Program Assessments by Impact Level and Treatment

This figure shows CDF plots of program assessments for control and treatment conditions by each of the four possible program impact levels. "1,000 Impact of Lowest Programs" refers to the set of programs that are 1,000 times more impactful than the lowest-impact combination for that program, and so forth. Assessments are presented in terms of the raw dollar values respondents selected rather than the scaled log values.

Table A.2: Order Effects: Sensitivity on Control Screens

	(1) Scaled Assessment
After IC X Scaled Impact	0.058
	(0.138)
After SS X Scaled Impact	-0.163
	(0.129)
After Impact Calculator	-0.583
	(0.654)
After Side-by-Side	0.519
	(0.603)
Scaled Impact	0.424***
	(0.097)
Observations	380
Median Assessment	$3 million
Program FE	Yes

This table plots sensitivity in the Control condition, by the order in which the control assessments appear. Sensitivity is no different when the Control condition appears after the treatment conditions, as indicated by the first three rows. Standard errors are in parentheses and are clustered at the respondent level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

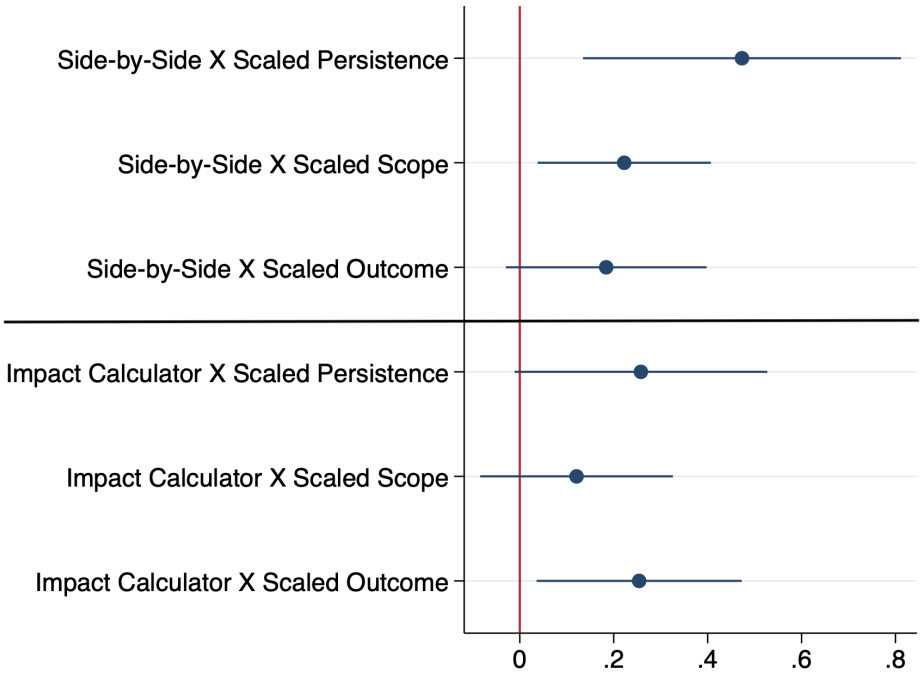


FIGURE A.5: Treatment Effects by Impact Feature

This figure plots the coefficients from a modified version of Equation 1, which interacts the three treatments included in the policymakers experiment with the log-scaled impact level for each impact feature. Error bars reflect 95% confidence intervals.

B General Public Sample - Additional Results

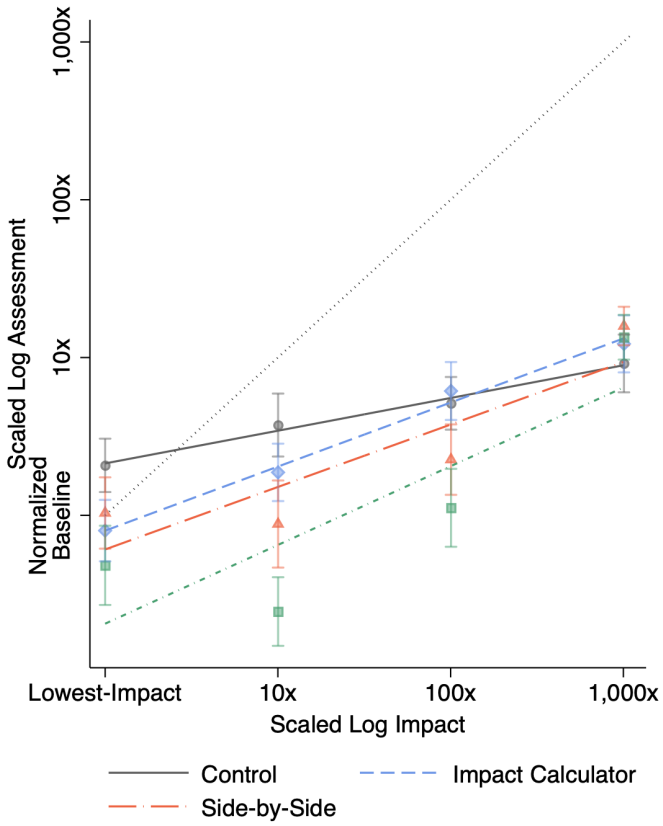


FIGURE B.1: Sensitivity Across General Public Study Conditions

This figure shows the relationship between program impact and the general public's assessments of program value in the control and three treatment conditions. The x-axis indicates the program impact, for each of the four possible impact combinations. The y-axis indicates the assessment of program value, compared to the average assessment provided for the lowest-impact variant of a program (computed separately for each program). Each point reflects the average program assessment for the corresponding impact level in a given condition, alongside 95% confidence intervals. Both impact and assessments are scaled according to the procedure described in Section 3.2.5.

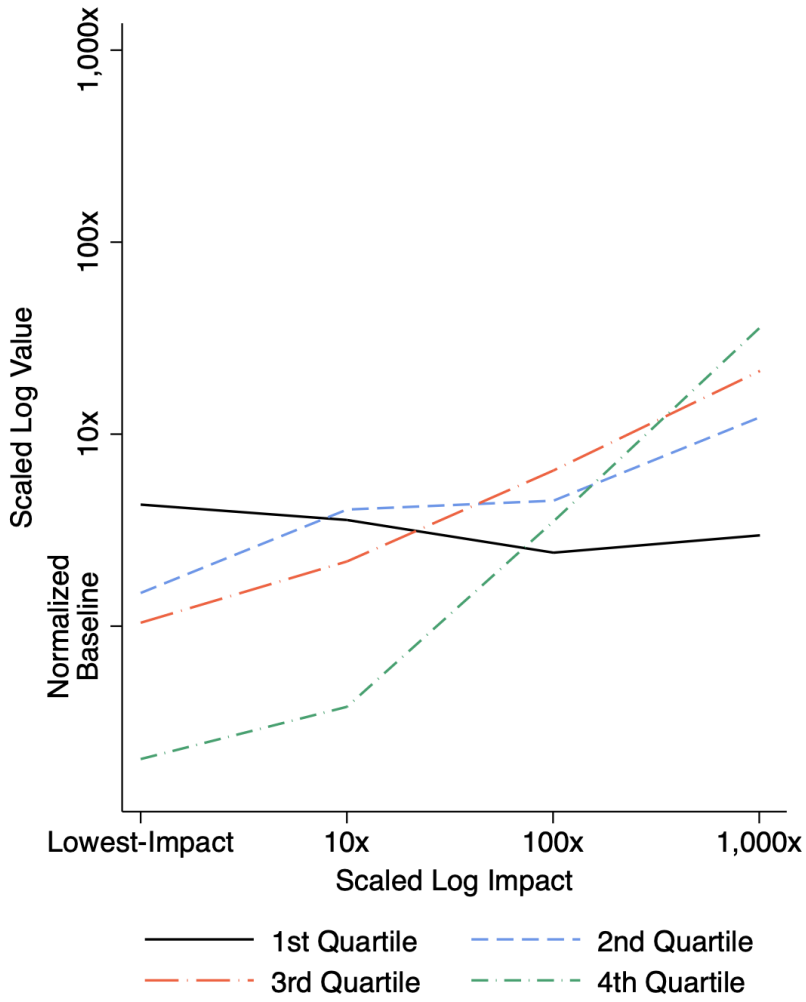


FIGURE B.2: Individual-Level Sensitivity Quartiles - General Public Sample

This figure plots the relationship between program assessments and program impact for respondents by quartile in terms of estimated sensitivity. Both impact and assessments are scaled according to the procedure described in Section 3.2.5.

Table B.1: Heterogeneities in Sensitivity - General Public Sample

	(1) Confidence	(2) Numeracy	(3) Conservative	(4) All
Factor X Scaled Impact	-0.001	0.249***	-0.121**	
	(0.001)	(0.071)	(0.052)	
Confidence	0.001			-0.000
	(0.006)			(0.001)
Numeracy (Perc Correct)		-0.752*		0.201***
		(0.426)		(0.074)
Lean Conservative			-0.340	-0.083
			(0.310)	(0.054)
Observations	2964	3000	3000	2964
Program FE	Yes	Yes	Yes	Yes
Scaled Impact	Yes	Yes	Yes	Yes
Treatment Controls	Yes	Yes	Yes	Yes

This table shows the relationship between key participant characteristics and sensitivity. The first row shows the coefficients of interest, i.e. the interaction between these characteristics and the scaled program impact. Certainty reflects self-reported certainty in program assessments, measured on a scale from 0 to 100. Numeracy is the percent of questions answered correctly on the four-question numeracy module, adapted from a longer numeracy assessment included in Kahan et al. (2012). Conservative is an indicator equal to one if the respondent indicates a value of 3 or above on a scale from 0 ('extremely liberal') to 7 ('extremely conservative'). The final column reports these interactions together in one regression. As in Table V, controls for program impact and treatment conditions are also included. Standard errors are in parentheses and are clustered at the respondent level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.2: Predicted Impact of Treatments on Sensitivity

	(1) Scaled Assessment	(2) Scaled Assessment
Pooled Treatment X Scaled Impact	0.195***	
	(0.052)	
Pooled Treatment	-0.949***	
	(0.235)	
Side-by-Side X Scaled Impact		0.264***
		(0.056)
Impact Calculator X Scaled Impact		0.165**
		(0.066)
Side-by-Side		-1.500***
		(0.294)
Impact Calculator		-0.614**
		(0.292)
Baseline Sensitivity	.16	.16
Respondent FE	Yes	Yes
Program FE	Yes	Yes
Impact Components	No	No
Observations	1500	1500

This table shows the results of an OLS regression relating *predicted* program assessments to program impact. Column 1 estimates the effect of being in any treatment condition, while Column 2 estimates the independent impact of each treatment. Both specifications control for sensitivity to the aggregated impact of a program derived from our three impact features, as described in Section 3.2.4. Standard errors are in parentheses and are clustered at the respondent level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

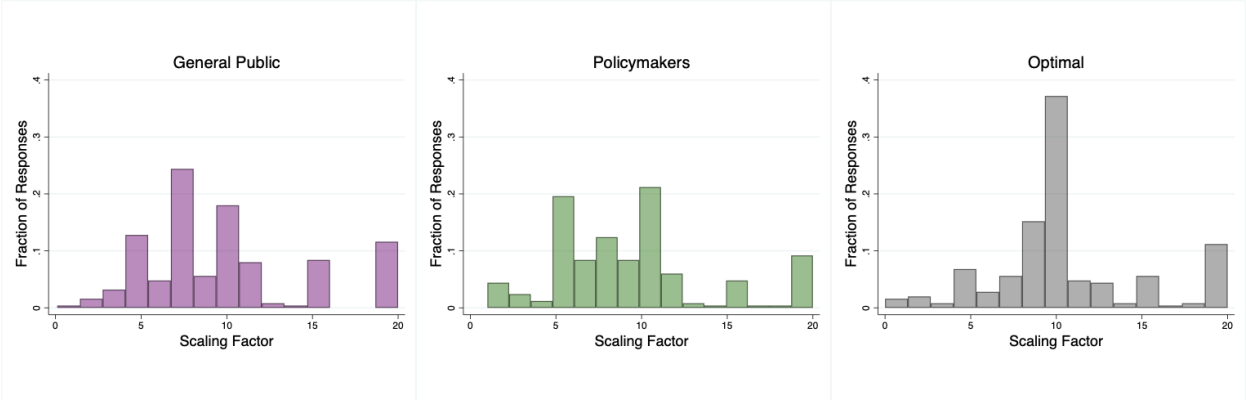


FIGURE B.3: Predicted Sensitivity by Population

This figure plots the frequency of responses to the following questions, asked of participants who made predictions about others' sensitivity: "By what factor do you think people's assessments of the value of a program change when its impact increases by a factor of 10?" Questions pertain to predictions of the general public sample, policymakers, and also what the participant herself thought the scaling factor "should" be. Responses are winsorized at a doubling of sensitivity, and 10 serves as the benchmark for one-to-one scaling between assessments and program impact.

C Multiple Hypothesis Corrections

We applied multiple hypothesis corrections across the treatment effects for the two decision aids, in accordance with the project's pre-analysis plan. The coefficients and adjusted p-values are reported in Table 4.2. To apply these corrections, we apply a bootstrap-based procedure to control the Family-Wise Error Rate (see for example List et al. (2019)). We took this approach rather than applying a formulaic correction (e.g. Holm or Bonferroni) in order to account for dependence among outcomes in our data. More specifically, our procedure followed the steps described below:

1. 5,000 bootstrap replications according to the following sub-steps:
   (a) Re-randomize the treatment and control conditions within individual.
   (b) Run the primary specification.
   (c) Save the t-stats computed for each regression coefficient, so the result is 5,000 t-stats for the impact of each decision aid.
2. Calculate the portion of cases in which the absolute value of at least one of the two bootstrapped t-stats are larger than each of the empirically-observed t-stats in turn. Each of these values reflects the probability of obtaining a result at least as extreme as our observed effect, in cases when the null hypothesis is true. These are our corrected p-values.

D Experiment Instructions - Policymakers Study

Respondents are high-ranking federal employees, recruited from across 22 U.S. government agencies, whose jobs involve developing and interpreting evidence and/or making adoption decisions based on program-relevant information. The experiment consists of six program assessments as well as follow-up questions to collect demographic information and learn more about evidence utilization within a respondent's own agency.

Figure D.1 shows the introduction screen.

FIGURE D.1: Introduction Screen

Figure D.2 provides instructions for the program assessments.

Program Assessments

In the first section of this survey you will **imagine you have the opportunity to help inform how different agencies allocate their budgets.**

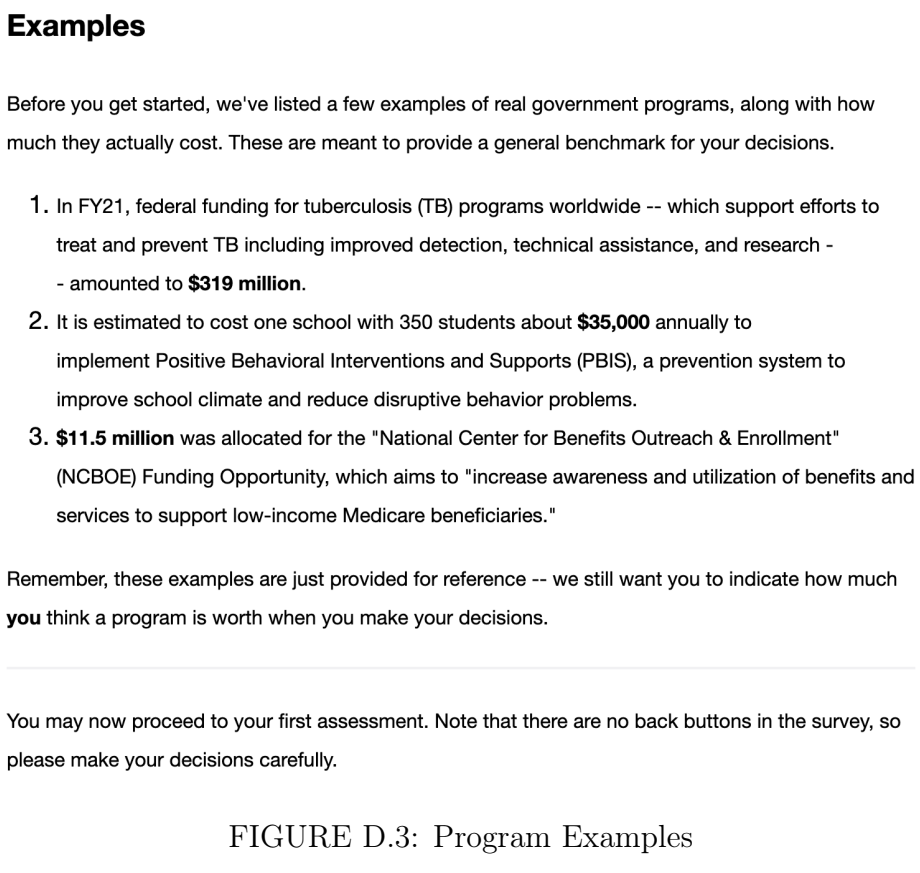
You will consider 5 hypothetical programs. For each program, you will indicate the maximum program cost such that you would still recommend the relevant agency fund the program.

Please note:

- Although all of these questions are hypothetical, please answer them as if the relevant agency would actually have to pay the cost of the program using the existing budget, and spend less on other priorities as a result.
- You will learn about the results of an evaluation of each hypothetical program. Information about how long program effects last can be interpreted as indicating the length of time the average program recipient showed positive outcomes compared to others who did not receive the program.
- When deciding whether you would support funding the program, **you should assume that the program would be implemented exactly as described in this survey.**

FIGURE D.2: Assessment Instructions

In Figure D.3, respondents are given examples of costs of actual government programs to benchmark their assessments.



Examples

Before you get started, we've listed a few examples of real government programs, along with how much they actually cost. These are meant to provide a general benchmark for your decisions.

1. In FY21, federal funding for tuberculosis (TB) programs worldwide -- which support efforts to treat and prevent TB including improved detection, technical assistance, and research -- amounted to **$319 million**.
2. It is estimated to cost one school with 350 students about **$35,000** annually to implement Positive Behavioral Interventions and Supports (PBIS), a prevention system to improve school climate and reduce disruptive behavior problems.
3. **$11.5 million** was allocated for the "National Center for Benefits Outreach & Enrollment" (NCBOE) Funding Opportunity, which aims to "increase awareness and utilization of benefits and services to support low-income Medicare beneficiaries."

Remember, these examples are just provided for reference -- we still want you to indicate how much **you** think a program is worth when you make your decisions.

You may now proceed to your first assessment. Note that there are no back buttons in the survey, so please make your decisions carefully.

FIGURE D.3: Program Examples

Figures D.4 and D.5 provide an example of a program description and decision screen in the Control condition. Respondents are asked to indicate the maximum amount the program could cost, such that they would recommend the program be funded at this cost but not for any higher cost listed. The control and treatment conditions appear in random order.

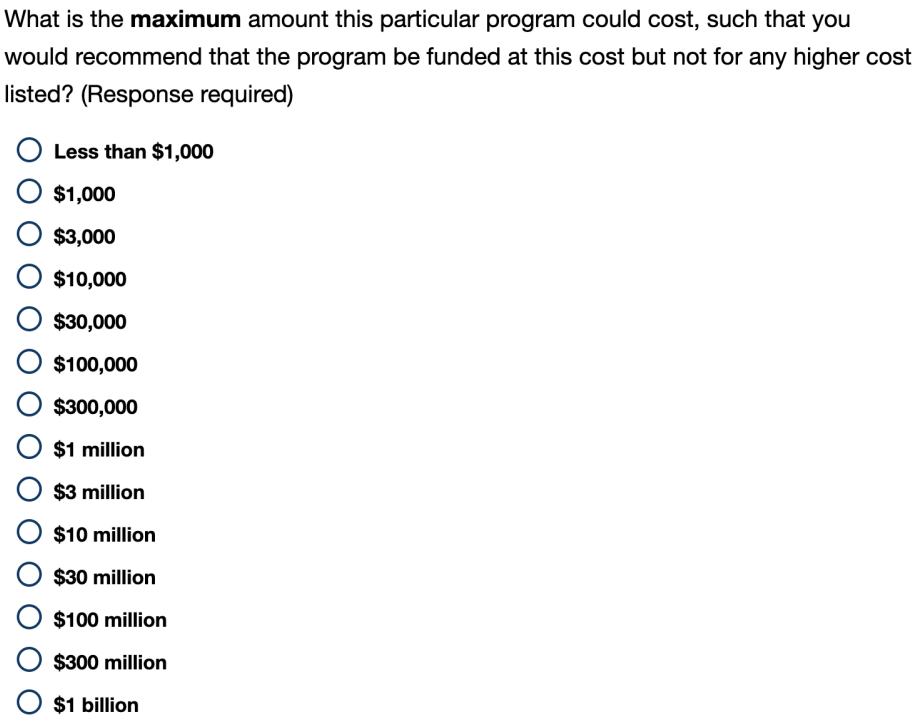
Program 2 out of 5

Consider a proactive outreach program designed to increase access to assistive technology for people with disabilities.

- **Number of people reached:** **600,000** people with disabilities.
- **Program outcome:** In an evaluation researchers found that the program increased **the likelihood of accessing assistive technology services** by 15 percentage points, over a baseline in which 47% of people with disabilities access assistive technologies.
- **How long effects last:** Effects persisted through all **5 years** of the evaluation.

Imagine you have the opportunity to help inform how different agencies allocate their budget.

FIGURE D.4: Control Program Screen



What is the **maximum** amount this particular program could cost, such that you would recommend that the program be funded at this cost but not for any higher cost listed? (Response required)

- ○ **Less than $1,000**
- ○ **$1,000**
- ○ **$3,000**
- ○ **$10,000**
- ○ **$30,000**
- ○ **$100,000**
- ○ **$300,000**
- ○ **$1 million**
- ○ **$3 million**
- ○ **$10 million**
- ○ **$30 million**
- ○ **$100 million**
- ○ **$300 million**
- ○ **$1 billion**

FIGURE D.5: Control Decision Screen

Figure D.6 shows the certainty elicitation, which appears after respondents make each of the two assessments in the Control condition. Respondents must indicate how certain they are that they gave the best possible assessment, on a scale from 0 to 100.

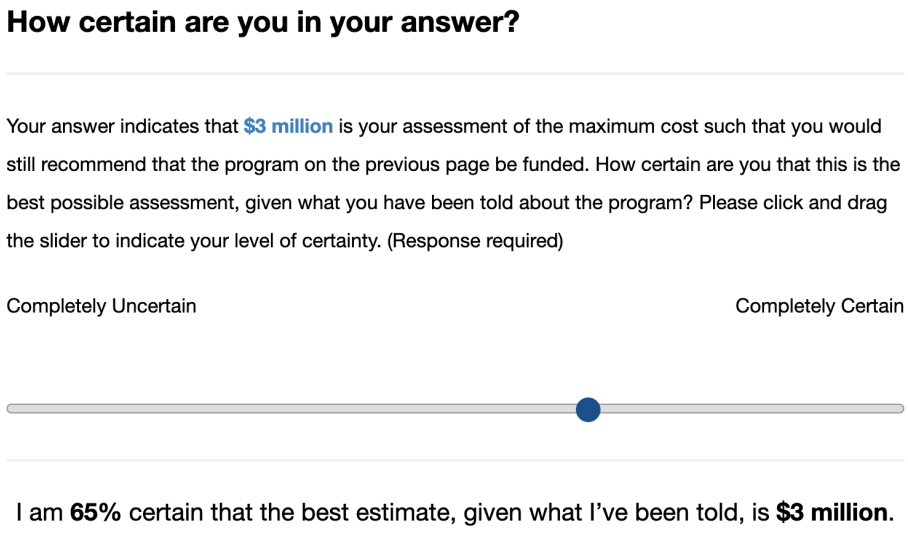
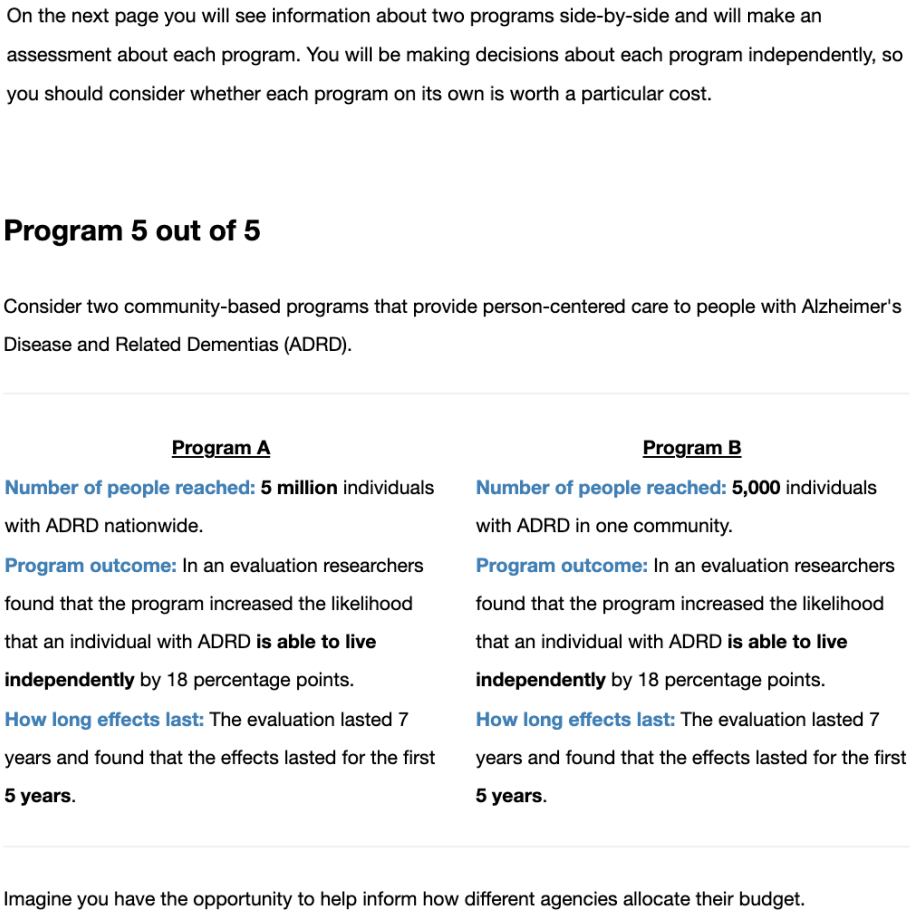


FIGURE D.6: Control Certainty Elicitation

Figure D.7 shows the Side-by-Side condition. Respondents are asked to make the same type of assessments as in the Control condition, but in this case two similar programs with different impact combinations appear together.



On the next page you will see information about two programs side-by-side and will make an assessment about each program. You will be making decisions about each program independently, so you should consider whether each program on its own is worth a particular cost.

Program 5 out of 5

Consider two community-based programs that provide person-centered care to people with Alzheimer's Disease and Related Dementias (ADRD).

Program A	Program B
Number of people reached: **5 million** individuals with ADRD nationwide.	**Number of people reached:** **5,000** individuals with ADRD in one community.
Program outcome: In an evaluation researchers found that the program increased the likelihood that an individual with ADRD **is able to live independently** by 18 percentage points.	**Program outcome:** In an evaluation researchers found that the program increased the likelihood that an individual with ADRD **is able to live independently** by 18 percentage points.
How long effects last: The evaluation lasted 7 years and found that the effects lasted for the first **5 years**.	**How long effects last:** The evaluation lasted 7 years and found that the effects lasted for the first **5 years**.

Imagine you have the opportunity to help inform how different agencies allocate their budget.

Imagine you have the opportunity to help inform how different agencies allocate their budget.

For each of the two programs, what is the **maximum** amount the program could cost, such that you would recommend that the program be funded at this cost but not for any higher cost listed? (Response required)

	Program A	Program B
Less than $1,000	○	○
$1,000	○	○
$3,000	○	○
$10,000	○	○
$30,000	○	○
$100,000	○	○
$300,000	○	○
$1 million	○	○
$3 million	○	○
$10 million	○	○
$30 million	○	○
$100 million	○	○
$300 million	○	○
$1 billion	○	○

FIGURE D.7: Side-by-Side Decision Screen

Figure D.8 shows the Impact Calculator condition. Respondents are asked to make the same type of assessments as in the Control condition, but in this case respondents additionally see a calculation of the number of people the program impacts per year.

On the following page, you will receive information about a hypothetical program. Before making your assessment you will also be shown an "impact calculator," which calculates the cost per additional person impacted by the program each year, given different possible total costs for the program.

Program 4 out of 5

Consider a proactive outreach program designed to increase take-up of TANF benefits. This program is in addition to the program that currently implements TANF.

- **Number of people reached:** **1 million** income-eligible families.
- **Program outcome:** In an evaluation researchers found that the program increased the likelihood of **clicking a link to the TANF website** by 9 percentage points. 1 in 1,000 individuals who clicked a link to the TANF website because of this program go on to take up TANF. At baseline, 25% of eligible families receive TANF benefits.
- **How long effects last:** The evaluation lasted 20 weeks and found that the effects lasted for **14 weeks**.

Imagine you have the opportunity to help inform how different agencies allocate their budget.

On this page, next to each possible total program cost we show in **blue** the **cost per additional person who takes up TANF cash assistance per year because of the program, who wouldn't have received TANF otherwise.** This is calculated based on each possible total program cost, and so as the proposed total cost increases, the cost per additional person who takes up TANF cash assistance per year increases proportionally.

What is the **maximum** amount this particular program could cost, such that you would recommend that the program be funded at this cost but not for any higher cost listed? (Response required)

◯ Total cost: **Less than $1,000**	Cost per additional person who takes up TANF cash assistance per year: **Less than $41.15**
◯ Total cost: **$1,000**	Cost per additional person who takes up TANF cash assistance per year: **$41.15**
◯ Total cost: **$3,000**	Cost per additional person who takes up TANF cash assistance per year: **$123.46**
◯ Total cost: **$10,000**	Cost per additional person who takes up TANF cash assistance per year: **$411.52**
◯ Total cost: **$30,000**	Cost per additional person who takes up TANF cash assistance per year: **$1,235**
◯ Total cost: **$100,000**	Cost per additional person who takes up TANF cash assistance per year: **$4,115**
◯ Total cost: **$300,000**	Cost per additional person who takes up TANF cash assistance per year: **$12,346**
◯ Total cost: **$1 million**	Cost per additional person who takes up TANF cash assistance per year: **$41,152**
◯ Total cost: **$3 million**	Cost per additional person who takes up TANF cash assistance per year: **$123,457**
◯ Total cost: **$10 million**	Cost per additional person who takes up TANF cash assistance per year: **$411,523**
◯ Total cost: **$30 million**	Cost per additional person who takes up TANF cash assistance per year: **$1.23 million**
◯ Total cost: **$100 million**	Cost per additional person who takes up TANF cash assistance per year: **$4.12 million**
◯ Total cost: **$300 million**	Cost per additional person who takes up TANF cash assistance per year: **$12.35 million**
◯ Total cost: **$1 billion**	Cost per additional person who takes up TANF cash assistance per year: **$41.15 million**

In case you're interested, we arrived at this number by multiplying the number of people impacted by the years the program is effective and then dividing the total cost by this amount, or:

$$\frac{\text{Cost}}{\text{Number Reached X Outcome X Years of Effect}}$$

FIGURE D.8: Impact Calculator Decision Screen

In Figure D.9, respondents estimate their relative certainty when making assessments on the “Impact Calculator" and “Side-by-Side” screens.

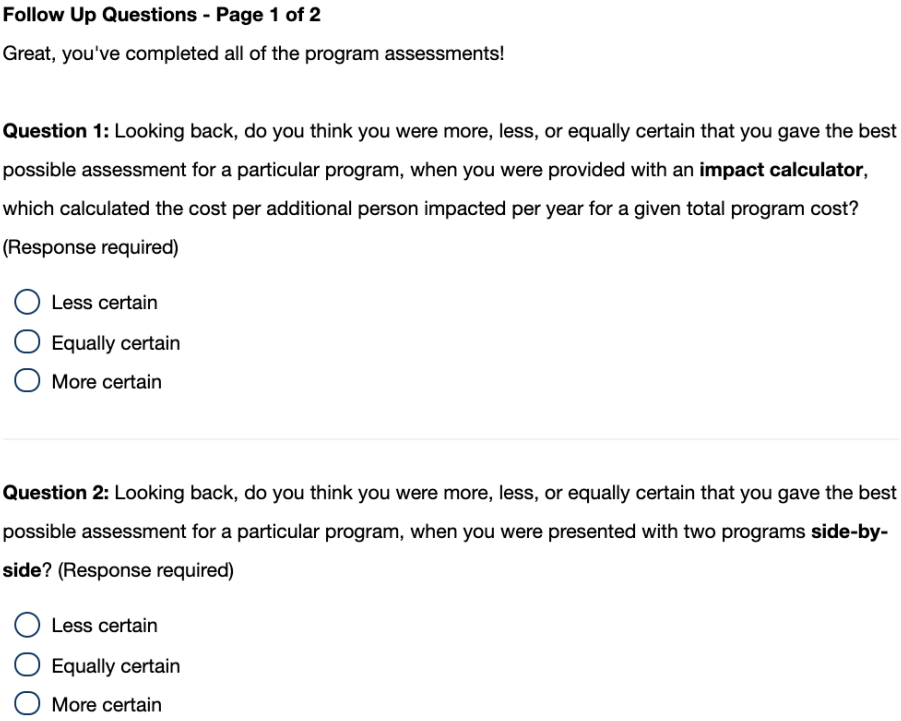
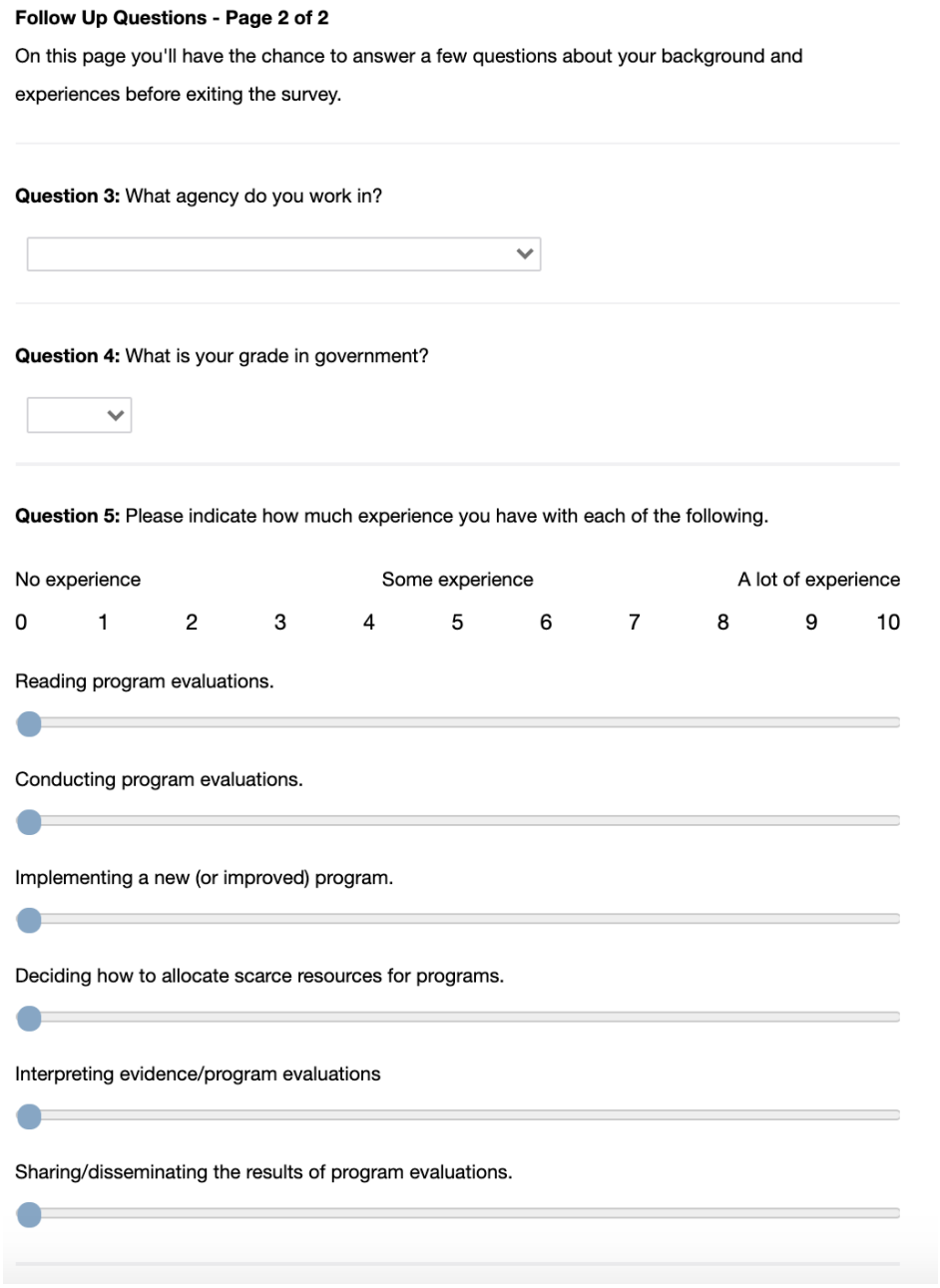


FIGURE D.9: Follow-Up Certainty Elicitations

Figure D.10 displays follow-up questions to conclude the survey. Questions 6 and 7 are randomly assigned to respondents. Questions vary slightly by survey type. For instance, some questions evaluating respondent experience ask additional questions specific to the agency at which the respondent works, while others do not ask about respondent grade in government when not applicable.



Question 6: If you could do one thing that would lead your agency to undertake more meaningful evaluations, what would it be?

Question 7: Which of the following barriers do you think interfere with the process of using evidence for programmatic decision making at your agency? (Check all that apply.)

- [ ] Need for technical support
- [ ] Not enough money in the budget
- [ ] Limited ability to influence program implementation
- [ ] Assumption that the existing programs do not need improvement
- [ ] Not enough time
- [ ] Preference to rely on own experience rather than evaluations
- [ ] Need for legal support
- [ ] Uncertainty about how to turn evidence into action
- [ ] Difficulty interpreting the implications of evidence-based recommendations
- [ ] Other, please specify:
- [ ] Existing evidence is not applicable to the program under consideration

Question 6: Please describe any program or program improvement you and/or your colleagues have been involved in implementing. You may skip this question if not applicable.

Short description of the program:	
Number of people reached:	
What kind of outcome the program addressed:	
How long effects last (if known):	
Estimated total cost to implement:	

Question 7: Please describe the most important factors that influence your agency's choice to fund or scale up a particular program or intervention.

Question 8: Please indicate the gender with which you most identify:

Male	Female	Non-binary / third gender	Prefer not to say
○	○	○	○

Question 9: Please indicate your age.

Question 10: Are you of Hispanic, Latino, or Spanish origin?

Yes	No	Prefer not to say
○	○	○

Question 11: How would you describe yourself?

- ☐ American Indian or Alaska Native
- ☐ Asian
- ☐ Black or African American
- ☐ Native Hawaiian or Other Pacific Islander
- ☐ White
- ☐ Prefer not to say
- ☐ Other:

Question 12: What is your highest level of education?

Less than a high school diploma	High school degree or equivalent (e.g. GED)	Some college, no degree	Associate degree (e.g. AA, AS)	Bachelor's degree (e.g. BA, BS)	Master's degree (e.g. MA, MS, MPP, MBA, MEd)	Professional degree (e.g. JD, MD, DDS, DVM)	Doctorate (e.g. PhD, EdD)	Prefer not to say
○	○	○	○	○	○	○	○	○

FIGURE D.10: Follow-Up Questions

E Recruitment materials

Recruitment materials vary based on recruitment method and respondent agency. Figure E.1 shows a representative invitation to participate.

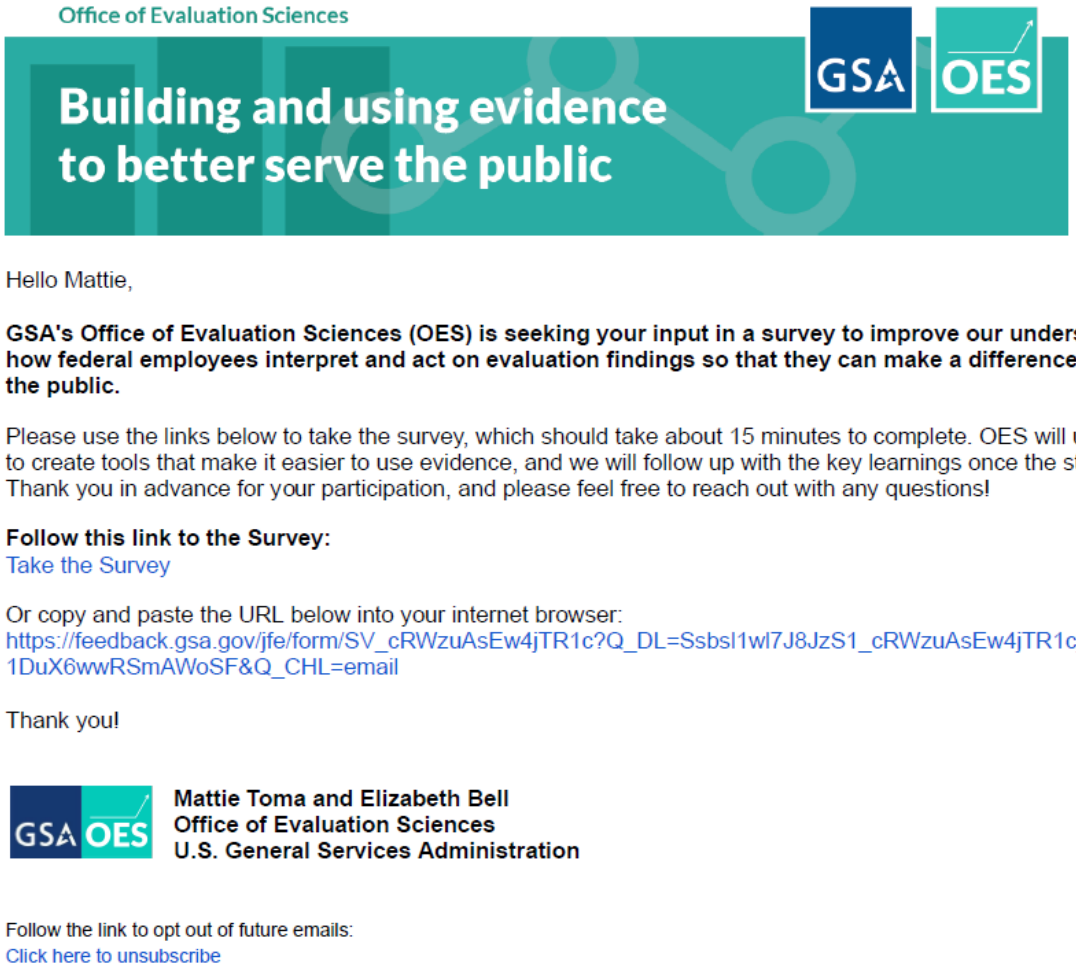


FIGURE E.1: Recruitment Materials

F Experimental Instructions - General Public Study

Participants are recruited via the online platform Prolific. Participation is restricted to individuals in the United States who had completed at least 100 studies with an overall approval rating of at least 95%. A representative sample is recruited based on the age, gender, and race breakdown reported in the 2019 U.S. Census American Community Survey data. Prior to participating in the study, participants must consent to participate.

Figure F.1 shows the introduction screen and comprehension question. 94% of respondents correctly answer this question.

Evaluations

In this survey, imagine you have the opportunity to help inform how different government agencies allocate their budgets. You will interpret evidence about 6 hypothetical policy programs and make assessments about the value of these programs. Specifically, you will be asked to indicate the maximum program cost such that you would recommend a government agency fund the program.

Please note:

- Although all of these questions are hypothetical, please answer them as if an agency would actually have to pay the cost of the program using the existing budget, and spend less on other priorities as a result.
- You will learn about the results of an evaluation of each hypothetical program. Information about how long program effects last can be interpreted as indicating the length of time the average program recipient showed positive outcomes compared to others who did not receive the program.
- When deciding whether you would support funding the program, **you should assume that the program would be implemented exactly as described in this survey.**

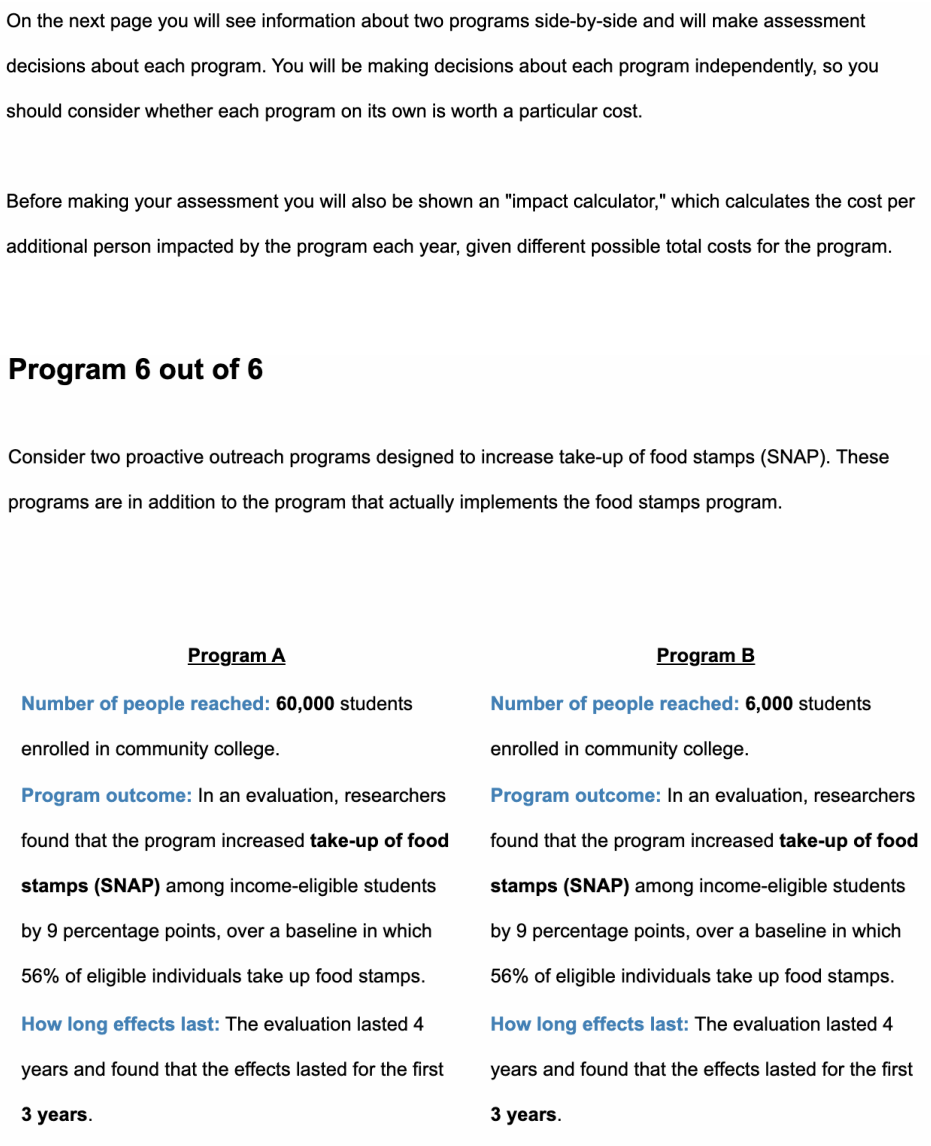
This survey is expected to take approximately 15 minutes to complete. After completing this study, you will receive a **$2.50 payment**, which will be distributed within 24 hours.

COMPREHENSION QUESTION: Imagine you said you would support funding a program for $1 million. What does this imply? Please select one of the following options.

- ◯ You support spending $1 million to run a pilot test of this program to see how well it actually works in practice.
- ◯ You support spending $1 million to roll this program out across all individuals in the US.
- ◯ You support spending $1 million to implement the program exactly as described in this survey.

FIGURE F.1: Introduction Screen

Figure F.2 shows the Joint Treatment condition. Respondents are asked to make the same type of assessments as in the Control condition, but in this case respondents see both the calculation of the number of people the program impacts per year as well as two similar programs with different impact combinations together on one screen.



On the next page you will see information about two programs side-by-side and will make assessment decisions about each program. You will be making decisions about each program independently, so you should consider whether each program on its own is worth a particular cost.

Before making your assessment you will also be shown an "impact calculator," which calculates the cost per additional person impacted by the program each year, given different possible total costs for the program.

Program 6 out of 6

Consider two proactive outreach programs designed to increase take-up of food stamps (SNAP). These programs are in addition to the program that actually implements the food stamps program.

Program A	Program B
Number of people reached: **60,000** students enrolled in community college.	**Number of people reached:** **6,000** students enrolled in community college.
Program outcome: In an evaluation, researchers found that the program increased **take-up of food stamps (SNAP)** among income-eligible students by 9 percentage points, over a baseline in which 56% of eligible individuals take up food stamps.	**Program outcome:** In an evaluation, researchers found that the program increased **take-up of food stamps (SNAP)** among income-eligible students by 9 percentage points, over a baseline in which 56% of eligible individuals take up food stamps.
How long effects last: The evaluation lasted 4 years and found that the effects lasted for the first **3 years**.	**How long effects last:** The evaluation lasted 4 years and found that the effects lasted for the first **3 years**.

Imagine you have the opportunity to help inform how an agency allocates its budget.

On this page, next to each possible total program cost we show in **blue** the **cost per additional student taking up SNAP per year, who wouldn't have taken up SNAP otherwise.** This is calculated based on each possible total program cost, and so as the proposed total cost increases, the cost per additional student taking up SNAP per year increases proportionally.

For each of the two programs, what is the **maximum** amount the program could cost, such that you would recommend that an agency fund the program at this cost but not for any higher cost listed?

Program A		Program B	
Total cost: **Less than $1,000**	Cost per additional student taking up SNAP per year: **Less than $0.06**	Total cost: **Less than $1,000**	Cost per additional student taking up SNAP per year: **Less than $0.62**
Total cost: **$1,000**	Cost per additional student taking up SNAP per year: **$0.06**	Total cost: **$1,000**	Cost per additional student taking up SNAP per year: **$0.62**
Total cost: **$3,000**	Cost per additional student taking up SNAP per year: **$0.19**	Total cost: **$3,000**	Cost per additional student taking up SNAP per year: **$1.85**
Total cost: **$10,000**	Cost per additional student taking up SNAP per year: **$0.62**	Total cost: **$10,000**	Cost per additional student taking up SNAP per year: **$6.17**
Total cost: **$30,000**	Cost per additional student taking up SNAP per year: **$1.85**	Total cost: **$30,000**	Cost per additional student taking up SNAP per year: **$18.52**
Total cost: **$100,000**	Cost per additional student taking up SNAP per year: **$6.17**	Total cost: **$100,000**	Cost per additional student taking up SNAP per year: **$61.73**

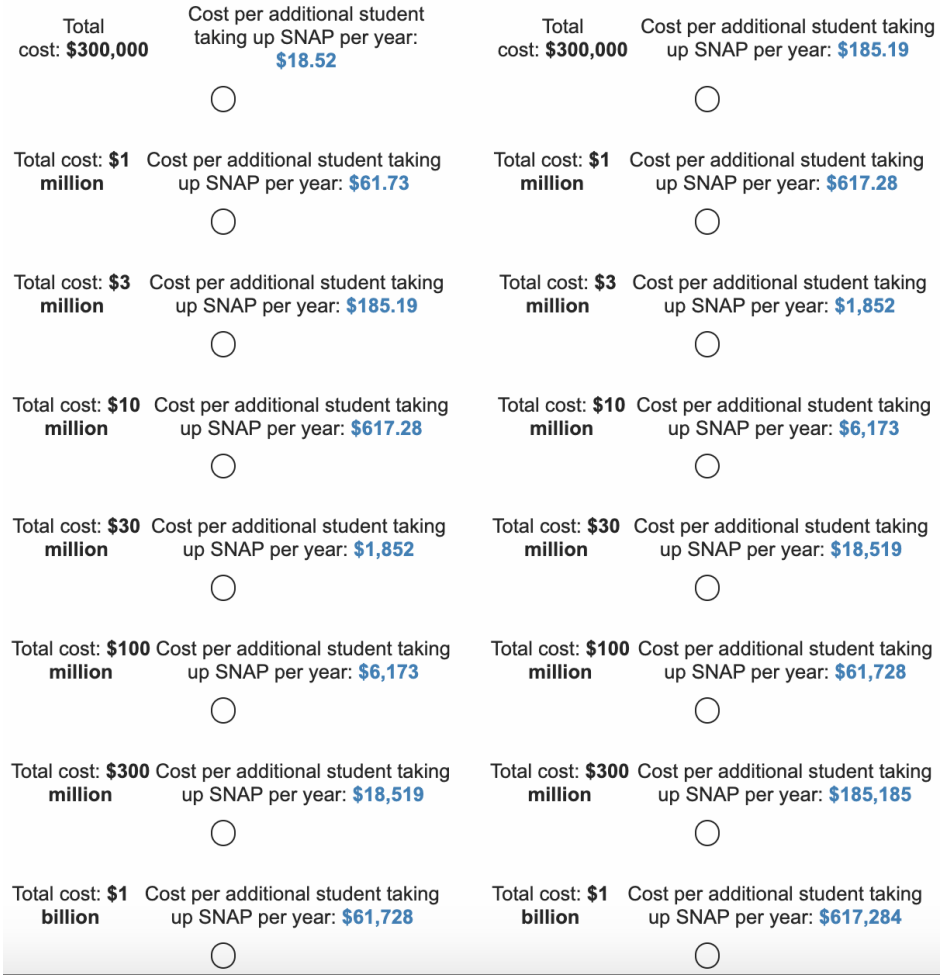
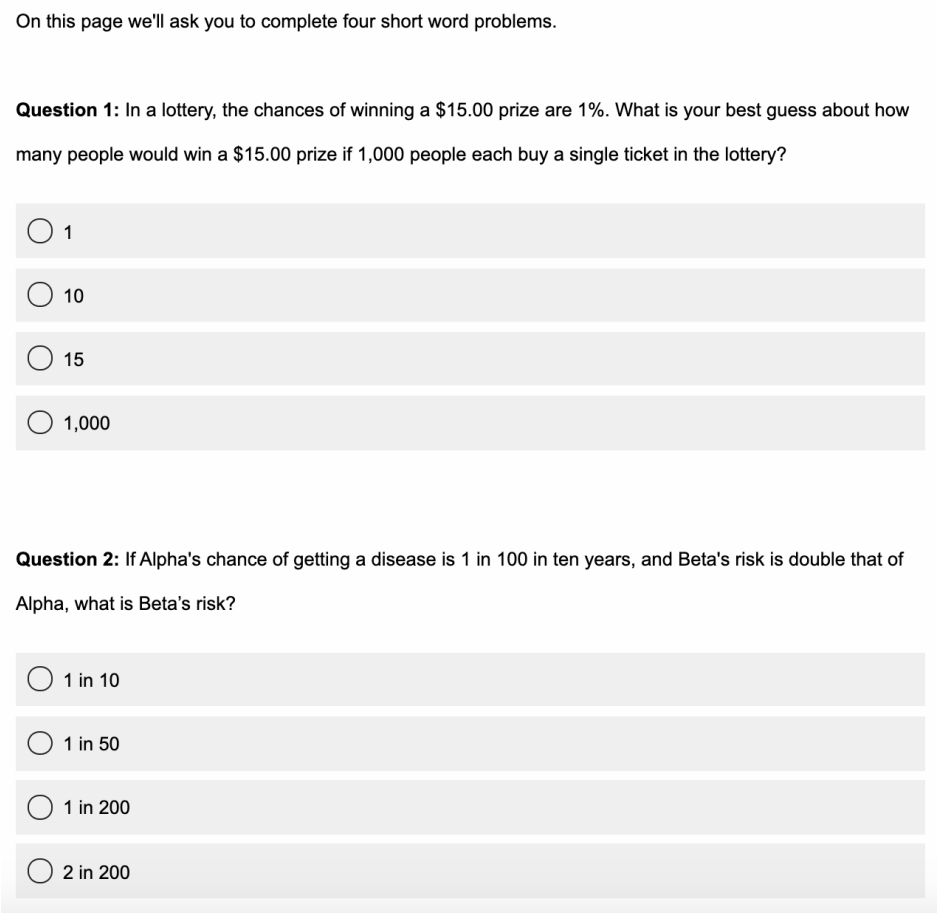


FIGURE F.2: Joint Treatment Decision Screen

Figure F.3 shows the four-question numeracy module, adapted from a longer numeracy assessment included in Kahan et al. (2012).


On this page we'll ask you to complete four short word problems.

Question 1: In a lottery, the chances of winning a $15.00 prize are 1%. What is your best guess about how many people would win a $15.00 prize if 1,000 people each buy a single ticket in the lottery?

- ◯ 1
- ◯ 10
- ◯ 15
- ◯ 1,000

Question 2: If Alpha's chance of getting a disease is 1 in 100 in ten years, and Beta's risk is double that of Alpha, what is Beta's risk?

- ◯ 1 in 10
- ◯ 1 in 50
- ◯ 1 in 200
- ◯ 2 in 200

Question 3: A mango and an orange cost $1.90 in total. The mango costs $1.00 more than the orange. How much does the orange cost?

- ◯ $0.45
- ◯ $0.90
- ◯ $1.00
- ◯ $1.45

Question 4: In a lake, there is a patch of lilypads. Every day, the patch doubles in size. If it takes 24 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

- ◯ 12 days
- ◯ 16 days
- ◯ 22 days
- ◯ 23 days

FIGURE F.3: Numeracy Questions

Figure F.4 displays follow-up questions to conclude the survey.

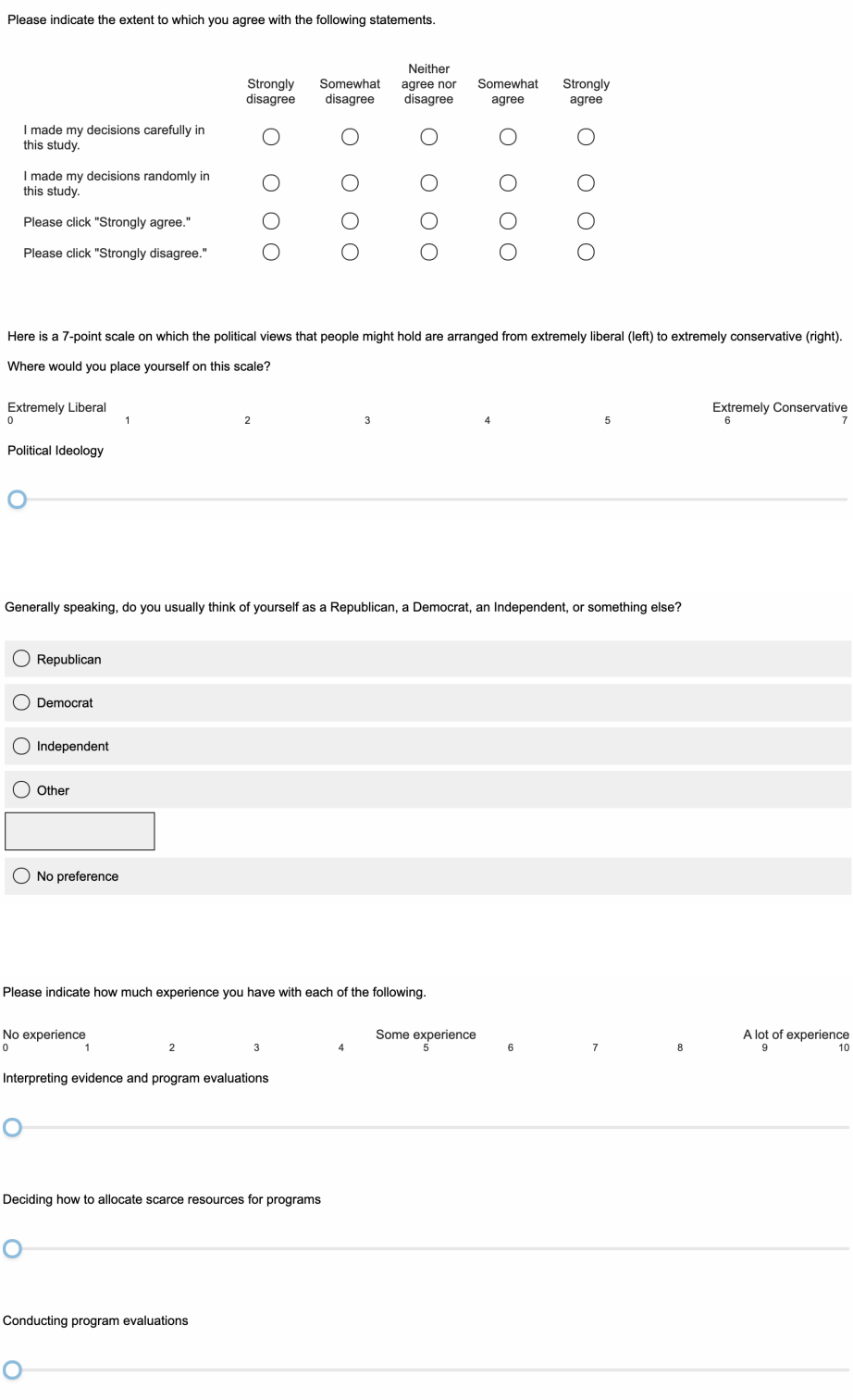


FIGURE F.4: Follow-Up Questions

Table G.1: Program Descriptions and Impact Combinations

Survey Type	Program Description	High Scope	Low Scope	High Outcome	Low Outcome	High Persistence	Low Persistence
Education	Consider a training program that provides first grade teachers with the tools to incorporate positive thinking techniques in math curricula.	2 million students nationwide	20,000 students in a school district	the program increased the likelihood that students passed national standardized math tests by 3 percentage points, over a baseline in which 34% of students passed	the program increased the likelihood of students adopting positive thinking skills according to a survey assessment by 3 percentage points. 1 in 1000 students who adopt positing thinking because of this program go on to pass national math standardized tests who wouldn't have otherwise. At baseline, 34% of students pass	Effects persisted through all 5 years of the evaluation	The evaluation lasted 5 years and found that the effects lasted for the first 6 months
Education	Consider a program that distributes guides encouraging families to regularly use a checklist to monitor their child's development.	100,000 families	10,000 families	the program led to a 4 percentage point increase in the likelihood that families actively utilized a developmental monitoring checklist , over a baseline in which 12% of families utilized a checklist	the program led to a 4 percentage point increase in the likelihood that families had access to a guide on developmental monitoring. 1 in 1000 families that access a guide because of this program go on to actively utilize a developmental monitoring checklist who wouldn't have otherwise. At baseline, 12% of families utilized a checklist	The evaluation lasted 5 years and found that the effects lasted for the first 4 years	The evaluation lasted 5 years and found that the effects lasted for the first 2 weeks
Education	Consider a program that provides child care assistance to single mothers enrolled in college with the goal of increasing degree attainment.	800,000 single mothers enrolled in college	800 single mothers enrolled in college	the program increased the likelihood of earning a college degree by 6 percentage points, over a baseline in which 46% receive a degree	the program increased the likelihood of continuous college enrollment with passing grades by 6 percentage points. 1 in 100 students who continue to enroll with passing grades because of this program go on to earn a degree who wouldn't have otherwise. At baseline, 46% of students receive a degree	Effects persisted through all 5 years of the evaluation	The evaluation lasted 5 years and found that the effects lasted for the first semester
Education	Consider a proactive outreach program designed to increase take-up of food stamps (SNAP). This program is in addition to the program that actually implements the food stamps program.	60,000 students enrolled in community college	6,000 students enrolled in community college	the program increased take-up of food stamps (SNAP) among income-eligible students by 9 percentage points, over a baseline in which 56% of eligible individuals take up food stamps	the program increased the likelihood that income-eligible students click a link to the website for the food stamps program (SNAP) by 9 percentage points. 1 in 100 students who clicked a link to the website because of this program go on to take up SNAP who wouldn't have otherwise. At baseline, 56% of eligible individuals take up food stamps	The evaluation lasted 4 years and found that the effects lasted for the first 3 years	The evaluation lasted 4 years and found that the effects lasted for the first day
Education	Consider a proactive outreach program for imprisoned individuals designed to increase take-up of the Pell Grant, which provides financial support to low-income students for college. This program is separate from the Pell Grant program itself.	1 million current or former prisoners	1,000 current or former prisoners	the outreach program increased the likelihood of receiving the Pell Grant by 4 percentage points, over a baseline in which 38% of eligible individuals receive the Pell Grant	the outreach program increased the likelihood of navigating to the website and creating a Federal Student Aid ID (FSA ID) by 4 percentage points. 1 in 10 individuals who create an FSA ID because of this program go on to take up the Pell Grant who wouldn't have otherwise. At baseline, 38% of eligible individuals receive the Pell Grant	The evaluation lasted 10 years and found that the effects lasted for the first 8 years	The evaluation lasted 10 years and found that the effects lasted for the first month
Health	Consider a proactive outreach program designed to increase take-up of TANF, the cash assistance program for people living in poverty. This program is in addition to the program that currently implements TANF.	1 million income-eligible families	100,000 income-eligible families	the program increased the likelihood of taking up TANF by 9 percentage points. At baseline, 25% of eligible families receive TANF benefits	the program increased the likelihood of clicking a link to the TANF website by 9 percentage points. 1 in 1,000 individuals who clicked a link to the TANF website because of this program go on to take up TANF. At baseline, 25% of eligible families receive TANF benefits	The evaluation lasted 20 weeks and found that the effects lasted for 14 weeks	The evaluation lasted 20 weeks and found that the effects lasted for the first day
Health	Consider a training program that provides preschool teachers in federally-funded Head Start schools with the tools to incorporate positive thinking techniques in math curricula.	2 million preschool students	2,000 preschool students	the program increased the likelihood that students passed national standardized math tests by 5 percentage points, over a baseline in which 34% of students passed	the program increased the likelihood that students adopted positive thinking skills by 5 percentage points. 1 in 100 students who adopt positive thinking because of this program go on to pass national math standardized tests who wouldn't have otherwise. At baseline, 34% of students pass	Effects persisted through all 10 years of the evaluation	The evaluation lasted 10 years and found that the effects lasted for the first year
Health	Consider a communications program designed to increase Head Start enrollment among refugee and migrant communities. This program is in addition to the program that currently implements Head Start, a federally-funded preschool program.	5 million refugee or migrant families	5,000 refugee or migrant families	the program increased the likelihood that families enrolled their child in Head Start by 10 percentage points. At baseline, 42% of eligible children were enrolled in Head Start	the program increased the likelihood that families navigated to the Head Start website by 10 percentage points. 1 in 100 families who navigated to the Head Start website because of this program go on to enroll their child in Head Start. At baseline, 42% of eligible children were enrolled in Head Start	Effects persisted through the full year of the evaluation	The evaluation lasted 1 year and found that the effects lasted for the first 5 weeks
Health	Consider a program that provides high-quality child care for low-income parents searching for a job.	1.5 million low-income parents searching for a job	15,000 low-income parents searching for a job	the program led to a 9 percentage point increase in the likelihood that low-income parents find a job	the program led to a 9 percentage point increase in the likelihood that low-income parents were able to apply to at least one job each day. 1 in 10 individuals who were able to apply to at least one job each day because of this program went on to find a job who wouldn't have otherwise	The evaluation lasted 4 years and found that the effects lasted for the first 3 years	The evaluation lasted 4 years and found that the effects lasted for the first day
Health	Consider a job training program for American Indian/Alaska Native people that provides skill development and work exposure.	100,000 American Indian/Alaska Native people	10,000 American Indian/Alaska Native people	the program increased the likelihood of finding a job by 14 percentage points	the program increased the likelihood of passing a job readiness assessment by 14 percentage points. 1 in 1,000 individuals who pass the job readiness assessment go on to find a job who wouldn't have otherwise	Effects persisted through all 5 years of the evaluation	The evaluation lasted 5 years and found that the effects lasted for the first 3 weeks

Survey Type	Program Description	High Scope	Low Scope	High Outcome	Low Outcome	High Persistence	Low Persistence
Health	Consider a proactive outreach program designed to increase access to assistive technology (i.e. wheelchairs, hearing aids, cognitive aids, etc.) for people with disabilities.	600,000 people with disabilities	6,000 people with disabilities	the program increased the likelihood of accessing assistive technology by 15 percentage points, over a baseline in which 47% of people with disabilities access assistive technologies	the program increased the likelihood of clicking a link to the State Assistive Technology Program Directory website by 15 percentage points. 1 in 10 people who clicked the link because of the program went on to access services who wouldn't have otherwise. At baseline, 47% of people with disabilities access assistive technologies	Effects persisted through all 5 years of the evaluation	The evaluation lasted 5 years and found that the effects lasted for the first 2 days
Health	Consider a public outreach program that provides caregivers with information on long-term services and support options for older adults in their care via personalized text messages.	1 million caregivers	100,000 caregivers	the outreach program increased the likelihood of caregivers accessing at least one long-term service or support by 24 percentage points. In the absence of the program, 30% of caregivers accessed at least one service or support	the outreach program increased self-reported awareness of long-term services and support options by 24 percentage points. 1 in 1000 people who were aware of these options because of the program went on to access at least one long-term service or support who wouldn't have otherwise. In the absence of the program, 30% of caregivers accessed at least one service or support	Effects persisted through all 5 years of the evaluation	The evaluation lasted 5 years and found that the effects lasted for the first 2 weeks
Health	Consider a program designed to provide assistance in lowering the costs of Medicare premiums and deductibles among people eligible for Medicare in tribal communities. Medicare is our country's health insurance program for people age 65 or older.	200,000 residents of tribal communities	200 residents of tribal communities	the program increased the likelihood of getting a lower Medicare premium and/or deductible by 9 percentage points. In the absence of the program, 11% of people make active coverage choices that reduce their premiums	the program increased the likelihood of self-reporting an intention to change plans by 9 percentage points. 1 in 100 people who self-reported an intention to change plans because of the program went on to get a lower Medicare premium and/or deductible who wouldn't have otherwise. In the absence of the program, 11% of people make active coverage choices that reduce their premiums	Effects persisted through all 10 years of the evaluation	The evaluation lasted 10 years and found that the effects lasted for the first year
Health	Consider a community-based program that provides person-centered care to people with Alzheimer's Disease and Related Dementias (ADRD), progressive brain disorders that slowly destroy memory and thinking skills.	5 million individuals with ADRD nationwide	5,000 individuals with ADRD in one community	the program increased the likelihood that an individual with ADRD is able to live independently by 18 percentage points	the program increased the likelihood that an individual with ADRD self-reports knowledge of key activities to manage their symptoms at home by 18 percentage points. 1 in 100 people who self-reported knowledge of ways to manage symptoms at home because of the program were able to live independently, who wouldn't have otherwise	The evaluation lasted 7 years and found that the effects lasted for the first 5 years	The evaluation lasted 7 years and found that the effects lasted for the first 6 months
Health	Consider an after school program that provides physical fitness and nutrition education for students with disabilities.	800,000 students with disabilities	80,000 students with disabilities	the program increased the likelihood of being in the Healthy Fitness Zone according to school fitness assessments by 15 percentage points, over a baseline in which 27% of students with disabilities are in the Healthy Fitness Zone	the program increased the likelihood of passing an assessment testing students' understanding of ways to improve physical well being by 15 percentage points. 1 in 1000 students with disabilities who passed because of the program went on to be in the Healthy Fitness Zone according to school fitness assessments who wouldn't have otherwise. At baseline, 27% of students with disabilities are in the Healthy Fitness Zone	The evaluation lasted 10 years and found that the effects lasted for the first 8 years	The evaluation lasted 10 years and found that the effects lasted for the first month
International Development	Consider a program designed to reduce the spread of misinformation and, in turn, promote efforts to support democracy, human rights, and good governance.	2 million individuals in post-transition countries	200,000 individuals in post-transition countries	the program decreased the likelihood of sharing misinformation on social media regularly by 6 percentage points, over a baseline in which 29% of users share misinformation at least once a week	the program increased the likelihood of correctly identifying misinformation in a survey by 6 percentage points. 1 in 1000 individuals who were able to correctly identify misinformation because of the program went on to stop sharing misinformation on social media regularly. At baseline, 29% of users share misinformation at least once a week	Effects persisted through all 5 years of the evaluation	The evaluation lasted 5 years and found that the effects lasted for the first 2 weeks
International Development	Consider an outreach program designed to increase childhood vaccination rates.	100,000 families in India	1,000 families in India	the program increased the likelihood that a child receives all recommended vaccinations by 9 percentage points, over a baseline in which 42% of children receive all required vaccinations	the program increased the likelihood that a family self-reports an intention to vaccinate their child by 9 percentage points. 1 in 10 families who self-reported an intention to vaccinate their child because of the program went on to get all recommended vaccines for their child who wouldn't have otherwise. At baseline, 42% of children receive all recommended vaccines	The evaluation lasted 10 years and found that the effects lasted for the first 5 years	The evaluation lasted 10 years and found that the effects lasted for the first 2 days
International Development	Consider a program designed to strengthen local government units via training and operational support with the goal of reducing crime.	25 million individuals in the Philippines	25,000 individuals in the Philippines	the program decreased the chance of any individual being a victim of violent crime by 0.5 percentage points. In the absence of the program, 4% of community members self-report at some point being a victim of a violent crime	the program increased the likelihood of utilizing a mediator to resolve a dispute by 0.5 percentage points. For every 10 disputes that use a mediator because of the program, 1 person who would have been a victim of violent crime otherwise did not become a victim. In the absence of the program, 4% of community members self-report at some point being a victim of a violent crime	The evaluation lasted 10 years and found that the effects lasted for the first 8 years	The evaluation lasted 10 years and found that the effects lasted for the first month
International Development	Consider a program that provides public teachers training on integrating technology into school curricula.	800,000 students in Kenya	8,000 students in Kenya	the likelihood that students mastered reading comprehension skills for their grade level increased by 3 percentage points, over a baseline in which 41% of students exhibited mastery	the likelihood that students used technology in the classroom increased by 3 percentage points. 1 in 1000 students who used technology in the classroom because of the program went on to master reading comprehension skills for their grade level who wouldn't have otherwise. At baseline, 41% of students exhibited mastery	Effects persisted through all 10 years of the evaluation	The evaluation lasted 10 years and found that the effects lasted for the first year
International Development	Consider a program that transfers cash directly to families in need with the goal of increasing long-run economic productivity.	400,000 households in Liberia	400 households in Liberia	the program increased the likelihood that a household's earnings were above $2 a day by 11 percentage points, over a baseline in which 55% of households earned more than $2 a day	the program increased the likelihood that a household invests in productive assests by 11 percentage points. 1 in 100 households that invested in productive assets because of the program saw their earnings go above $2 a day who wouldn't have otherwise. At baseline, 55% of households earned more than $2 a day	Effects persisted through all 2.5 years of the evaluation	The evaluation lasted 2.5 years and found that the effects lasted for the first 3 months