Analysis Plan

Project Name: Understanding and Improving Policymakers' Sensitivity to Impact

Project Code: 2101

Date Finalized: 05/21/2021

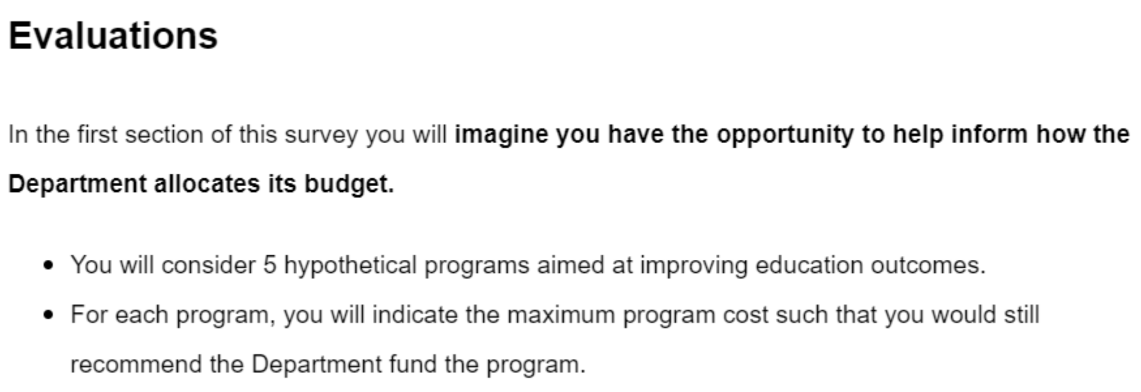
Project Description

This project leverages an online survey experiment among federal employees involved in the process of translating information and evidence about a program into actionable policies and programs. The experiment's primary aim is to test ways of presenting information to improve *sensitivity* to impact-relevant information about a program. In addition, the experiment seeks to improve our understanding of what the process of making decisions about programs and policies looks like across the federal agencies involved in the experiment.

Federal employees will be invited to take part in a short online survey via the General Services Administration Qualtrics platform.

Participants will see five hypothetical program descriptions and will provide a **"valuation"**, defined as an assessment of the maximum cost at which they think it would be worth funding the program.

The prompt respondents see before making these decisions is as follows. Note, all survey extracts in this document use a version of the survey created for the Department of Education:



Evaluations

In the first section of this survey you will **imagine you have the opportunity to help inform how the Department allocates its budget.**

- You will consider 5 hypothetical programs aimed at improving education outcomes.
- For each program, you will indicate the maximum program cost such that you would still recommend the Department fund the program.

Each hypothetical program description will include randomly-assigned variations corresponding to different hypothesized impact-relevant features of a program (henceforth called **"impact features"**).

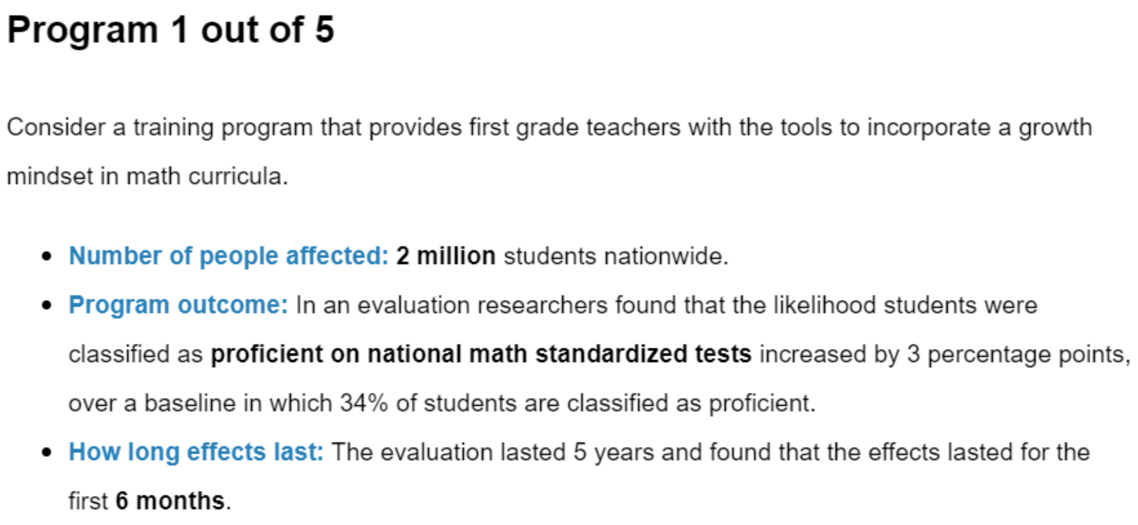
- ***Number of people impacted*** - How does valuation depend on the scope of a program?
- ***Intermediate versus final outcomes*** - How does valuation depend on whether the intervention impacts an intermediate outcome (e.g. click rates on an ad for a health program) versus a final outcome (e.g. enrollment, or even health outcomes)?
- ***Persistence*** - How does valuation depend on whether effects are documented for a shorter or longer amount of time?

These features can be multiplied together to estimate the impact of a program. Specifically, multiplying the number of people a program affects, the rate at which a "final" outcome is achieved, and the persistence of effects gives us an estimate for the *people affected per year*, who wouldn't achieve the outcome in question in the absence of the program. Each respondent will see one of four combinations of these variations for each program:

1. High scope, Final outcome, Long persistence
2. Low scope, Final outcome, Long persistence
3. High scope, Intermediate outcome, Long persistence
4. High scope, Final outcome, Short persistence

We have constructed the features such that the four combinations will be 1x, 10x, 100x, and 1000x the lowest impact level.

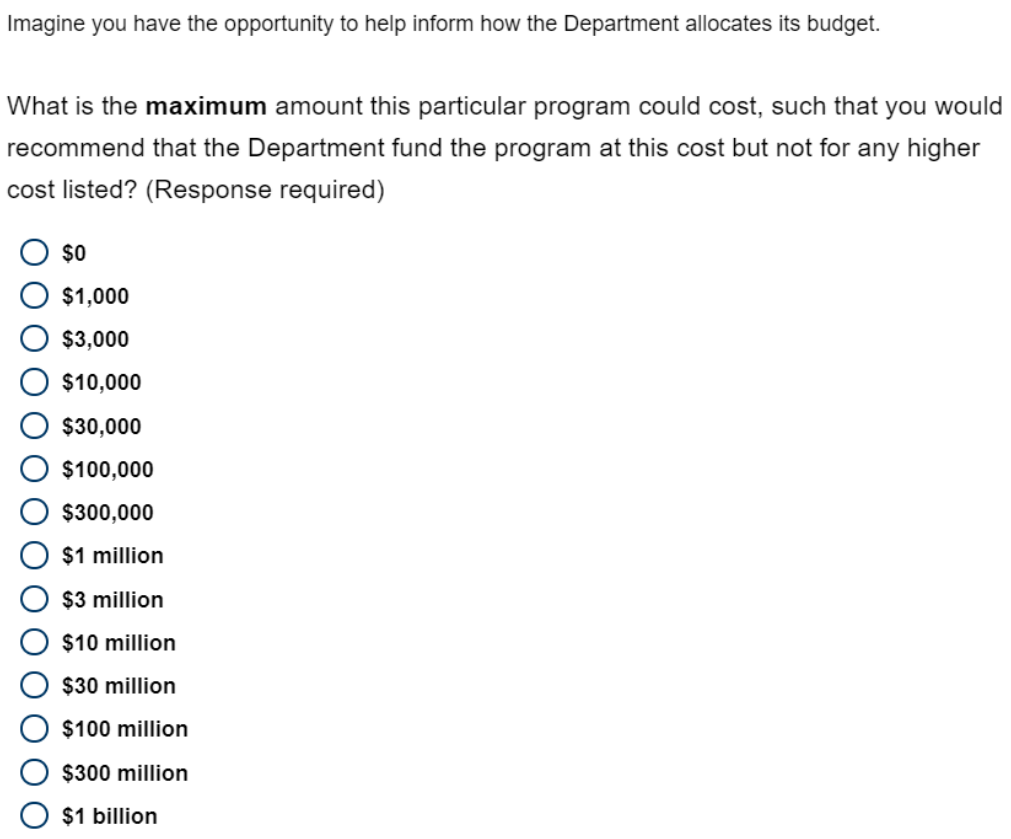
An example of a program description with randomly-assigned features is as follows:



Program 1 out of 5

Consider a training program that provides first grade teachers with the tools to incorporate a growth mindset in math curricula.

- **Number of people affected:** **2 million** students nationwide.
- **Program outcome:** In an evaluation researchers found that the likelihood students were classified as **proficient on national math standardized tests** increased by 3 percentage points, over a baseline in which 34% of students are classified as proficient.
- **How long effects last:** The evaluation lasted 5 years and found that the effects lasted for the first **6 months**.



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

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

○ $0
○ $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

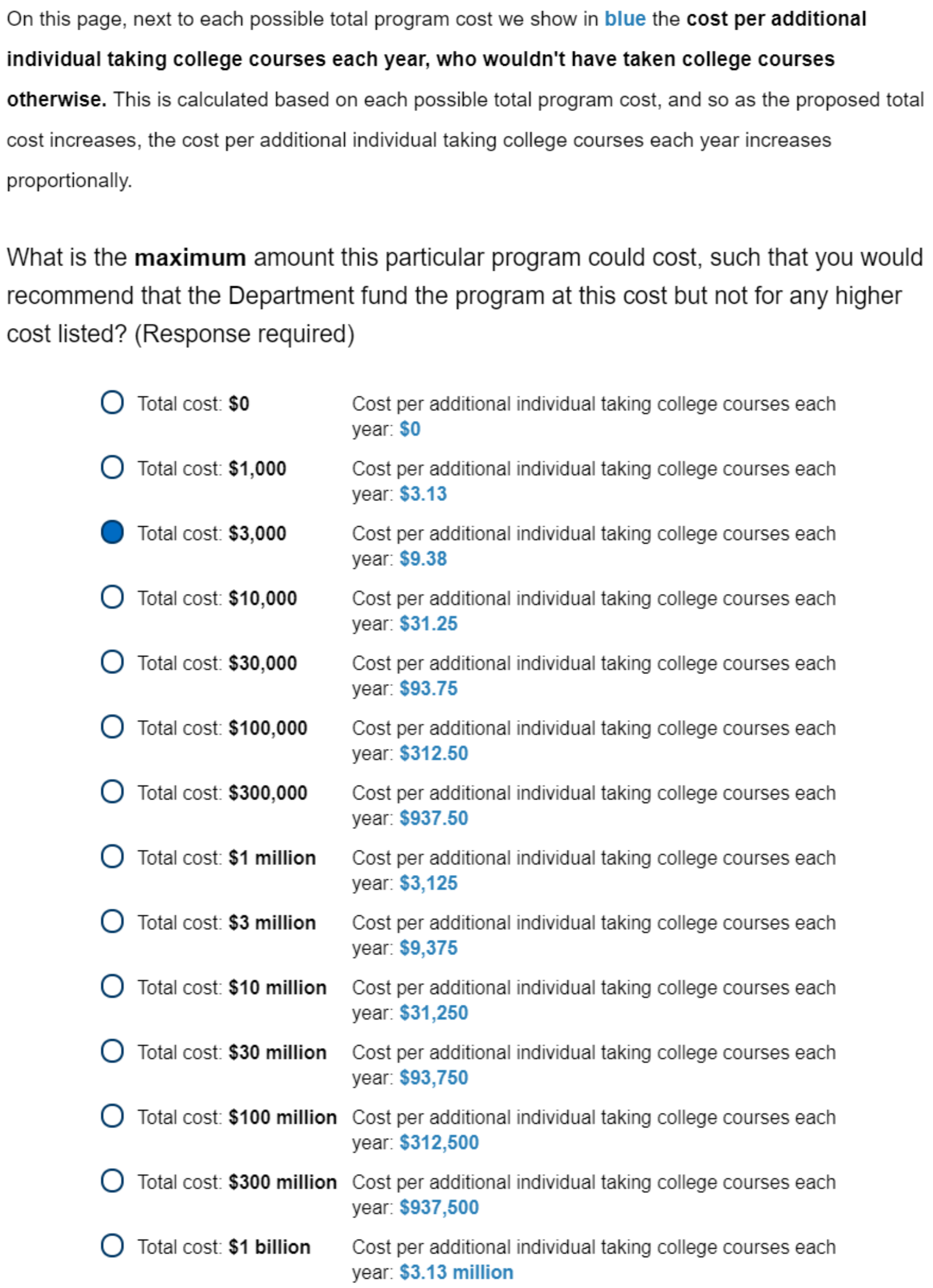
The survey will be broken down into three sections, which each respondent sees in random order:

1. **Baseline - 2 valuations, 2 unique programs:** Document baseline insensitivities in "valuations"; to what extent are policymakers attending to features like scope or outcome type when assessing a program? The baseline questions will look like the sample decision screen shown above. Refer to the Outcomes section for additional information on how we are operationalizing sensitivities.

2. **Treatment 1 - Impact Calculator - 2 valuations, 2 unique programs:** Introduce an "Impact calculator" to increase sensitivity to impact-relevant features of a program. If respondents are presented with a calculation of impact (based only on the program description on the page), does sensitivity increase? If so, this would suggest value in adding impact metrics to OES abstracts or other dissemination materials.

3. **Treatment 2 - Side-by-Side - 2 valuations, 1 unique program:** Show program descriptions side-by-side rather than sequentially. Does the direct comparison increase sensitivity to impact-relevant features? If so, this would suggest value in presenting a new program alongside information on the status quo program when disseminating information. It would also recommend joint evaluations when decisions are being made.

An example screen for the impact calculation is shown here:

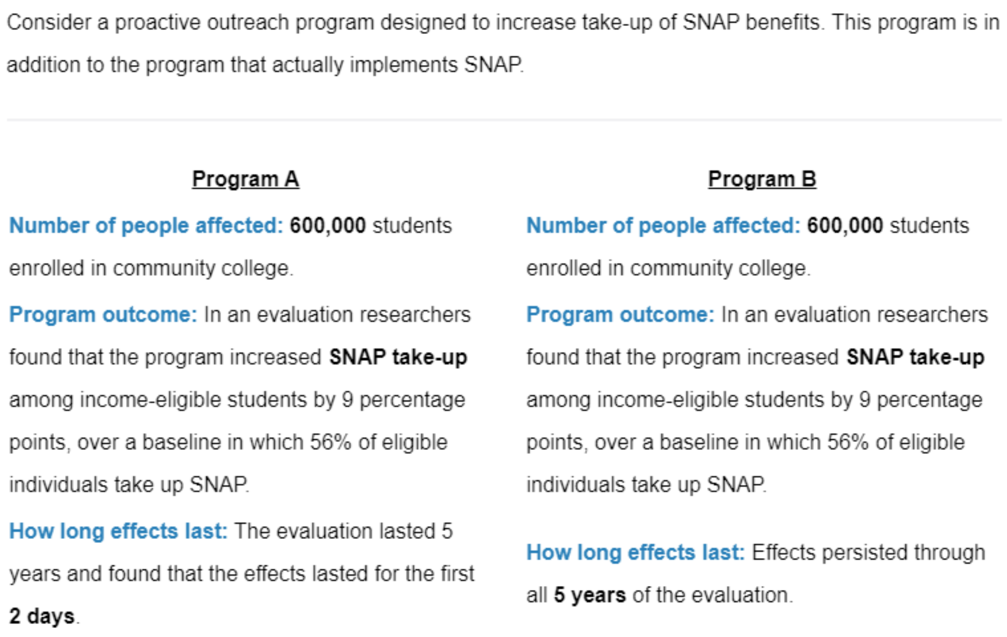


On this page, next to each possible total program cost we show in **blue** the **cost per additional individual taking college courses each year, who wouldn't have taken college courses otherwise.** This is calculated based on each possible total program cost, and so as the proposed total cost increases,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

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

○ Total cost: **$0**	Cost per additional individual taking college courses each year: **$0**
○ Total cost: **$1,000**	Cost per additional individual taking college courses each year: **$3.13**
● Total cost: **$3,000**	Cost per additional individual taking college courses each year: **$9.38**
○ Total cost: **$10,000**	Cost per additional individual taking college courses each year: **$31.25**
○ Total cost: **$30,000**	Cost per additional individual taking college courses each year: **$93.75**
○ Total cost: **$100,000**	Cost per additional individual taking college courses each year: **$312.50**
○ Total cost: **$300,000**	Cost per additional individual taking college courses each year: **$937.50**
○ Total cost: **$1 million**	Cost per additional individual taking college courses each year: **$3,125**
○ Total cost: **$3 million**	Cost per additional individual taking college courses each year: **$9,375**
○ Total cost: **$10 million**	Cost per additional individual taking college courses each year: **$31,250**
○ Total cost: **$30 million**	Cost per additional individual taking college courses each year: **$93,750**
○ Total cost: **$100 million**	Cost per additional individual taking college courses each year: **$312,500**
○ Total cost: **$300 million**	Cost per additional individual taking college courses each year: **$937,500**
○ Total cost: **$1 billion**	Cost per additional individual taking college courses each year: **$3.13 million**

An example screen for the side-by-side comparison is shown here:

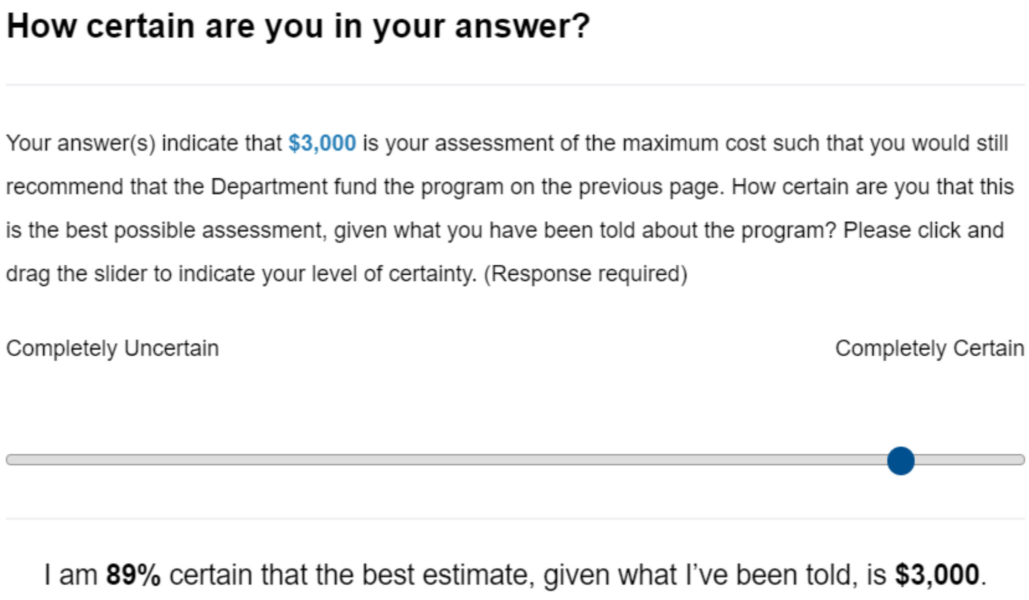


Consider a proactive outreach program designed to increase take-up of SNAP benefits. This program is in addition to the program that actually implements SNAP.

Program A	**Program B**
Number of people affected: **600,000** students enrolled in community college.	**Number of people affected:** **600,000** students enrolled in community college.
Program outcome: In an evaluation researchers found that the program increased **SNAP take-up** among income-eligible students by 9 percentage points, over a baseline in which 56% of eligible individuals take up SNAP.	**Program outcome:** In an evaluation researchers found that the program increased **SNAP take-up** among income-eligible students by 9 percentage points, over a baseline in which 56% of eligible individuals take up SNAP.
How long effects last: The evaluation lasted 5 years and found that the effects lasted for the first **2 days**.	**How long effects last:** Effects persisted through all **5 years** of the evaluation.

The three treatments allow us to document the degree to which decision makers are (in)sensitive to impact-relevant features of a program (Baseline), as well as whether particular modes of presenting evidence (Impact Calculator and Side-by-Side) can improve sensitivities (**our primary question of interest**).

After respondents see each baseline question, they will then be asked to assess their certainty in their valuation. Specifically, they will answer the following question:



How certain are you in your answer?

Your answer(s) indicate that **$3,000** is your assessment of the maximum cost such that you would still recommend that the Department fund the program on the previous page. How certain are you that this is the best possible assessment, given what you have been told about the program? Please click and drag the slider to indicate your level of certainty. (Response required)

I am **89%** certain that the best estimate, given what I've been told, is **$3,000**.

These questions will also allow for a heterogeneity analysis at baseline; for instance, are the decision makers who exhibit the most insensitivity in their valuation assessments or who are least familiar with evidence-based evaluation or program adoption decisions those with the most uncertainty? (Or, the converse may also be true.) Additional certainty questions asked at the end of the survey will also allow us to test whether modes of presenting information to improve sensitivity decrease decision makers' uncertainty in their assessments.

After responding to these six core questions, respondents will again be shown one of the baseline program descriptions. This time, they will be asked to predict the modal cost selected by *other respondents*. Because this beliefs-based question has a clear true or false answer, we are able to incentivize responses.

The survey will conclude with questions to understand the additional barriers policymakers confront when translating evidence into policies and programs, as well as what this process looks like at different agencies. The survey will also ask for information about respondents' work in government as well as demographic characteristics.

Data and Data Structure

This section describes variables that will be analyzed, as well as changes that will be made to the raw data with respect to data structure and variables.

Data Source(s):
All data will be collected and stored in the GSA Qualtrics platform.

Outcome Variables to Be Analyzed:

Primary Analysis:
The core test of interest is whether the treatments -- the Impact Calculator and Side-by-Side presentation -- improve sensitivity to impact-relevant features of evidence. Note that the particular specifications to identify treatment effects are included in the Statistical Models section below.

The primary outcome for this analysis is the ***"valuation" or perceived program value***, which is defined as the maximum cost at which the respondent is willing to fund the program, as identified in the survey.

We log-transform this outcome given that the costs are presented on a semi-logarithmic scale. We then normalize the program valuations around the mean perceived value ascribed to the lowest-impact combination for the program. This allows us to account for differences in average valuations and also presents a scaling factor that maps onto the measure of impact (described below in the Transformations of Variables section).

Secondary Analysis:

Secondary questions of interest include the following:

1. How does sensitivity and the impact of our treatments depend on the particular impact-relevant feature? For instance, are people differentially sensitive to a change in scope compared to a change in the outcome?
2. Is higher certainty correlated with increased sensitivity to impact?
3. Do the impact calculator and/or side-by-side presentations increase certainty in responses?
4. Does measured sensitivity look comparable across incentivized and unincentivized responses?
5. Is more experience with evidence and evaluation correlated with both increased sensitivity as well as higher certainty?
6. Is the time spent on the program valuation screens predictive of increased sensitivity?

The only new outcome measure these tests introduce is that of ***certainty*** when exploring heterogeneities. The baseline certainty measure will be coded as a continuous variable ranging from 0% to 100% certainty (or certain). The certainty measure relevant to treatment effects will be coded as -1 if respondents report less certainty in the face of a treatment, 0 if respondents are equally certain, and 1 if respondents report more certainty.

Transformations of Variables:

Note that the transformation for the main outcome of interest, **"valuation" or perceived program value,** is described in the Outcomes section above.

Program Impact: As noted above, the impact of each program (with varied impact features) can be calculated by multiplying the scope, outcome, and persistence of effects. This gives us a value for the people affected per year. We will capture the relative impact across the combinations of different impact features, such that we will code the impact as 1, 10, 100, or 1000 depending on whether we are looking at the lowest impact level for the program, the impact level that is 10x the baseline, and so forth. This value will be log-transformed and will serve as a key independent variable.

Program Features Impact: We will create variables equal to the scaled log program impact, as defined above, for each of our three impact features: scope, outcome, and persistence. For instance, the impact level for scope will be = 0 when the program description in question shows the lowest impact value for scope, log(10) at the next impact level, and so forth. These variables will indicate how the particular feature scales compared to its baseline value.

Treatment Indicators: We will create two dummy variables, *impact calculator* and *side-by-side*, that will equal one if the specified mode of presenting evidence is used for the particular question and zero otherwise. We will also create an index indicator, *treatment*, which will equal one if either the impact calculator or side-by-side mode is used, and zero otherwise.

Experience with evidence and evaluation: We will create an index for experience with evidence and evaluation based on our six-item scale at the end of the survey (variable name=experience). We will add together each individual item and divide by the total number of items (six) to get an average score for each respondent that can be used for heterogeneity analysis.

Time spent: We will compute the time spent, in seconds, on each program valuation screen.

Imported Variables:
NA

Transformations of Data Structure:
We will receive data at the respondent level and will reshape it such that it is at the decision screen level. That is, the final dataset will have six observations per respondent, one for each costing assessment they made.

Data Exclusion:
We will exclude respondents who didn't complete the main portion of the survey. We will also only keep the first complete response received via a single personalized survey link. That is, if the survey was taken multiple times using the same survey ID, we will not look at additional complete responses. We will also run robustness checks where we look only at respondents who spent at least 4 minutes on the survey in total.

Treatment of Missing Data:
We will require responses to all core program valuation and certainty questions. Any skipped end-of-survey questions will be dropped, without imputing the data.

Descriptive Statistics, Tables, & Graphs

The core graph we will present in both the OES abstract and an academic paper will show the scaled log impact on the x axis and the scaled log program valuations on the y axis, and will plot values across the baseline and treatment conditions. We hypothesize that valuations will be more responsive (sensitive) to a change in the impact for treatment compared to baseline decision screens.

Statistical Models & Hypothesis Tests

This section describes the statistical models and hypothesis tests that will make up the analysis — including any follow-ups on effects in the main statistical model and any exploratory analyses that can be anticipated prior to analysis.

Statistical Models:

Primary Analysis:

The primary specification will look at the impact of the two treatments (impact calculator and side-by-side) on increasing sensitivity to impact-relevant features of a program:

$$Y_{ip} = \beta_0 + \beta_1 IC_{ip} \cdot I_{ip} + \beta_2 SS_{ip} \cdot I_{ip} + \beta_3 IC_{ip} + \beta_4 SS_{ip} + \beta_5 I_{ip} + \delta_i + \alpha_p + \varepsilon_{ip}$$

where i indexes respondents evaluating program type p:

- Y_{ip} is our primary outcome of interest, i.e. the scaled log("valuation" or perceived program value) as defined above
- IC_{ip} is an indicator equal to one if the respondent sees an impact calculator on the screen when assessing the program
- SS_{ip} is an indicator equal to one if the respondent sees a side-by-side comparison of two similar programs when assessing the program
- I_{ip} is the scaled log(program impact), as defined above
- δ_i captures respondent fixed effects
- α_p captures program fixed effects

Robust standard errors will be adjusted to reflect clustering at the respondent level.

In this specification, our main tests of interest are whether:

- $\beta_1 > 0$, or that the impact calculator increases sensitivity to a change in impact
- $\beta_2 > 0$, or that the side-by-side presentation increases sensitivity to a change in impact

We will also look at an index that looks at the joint effects of the impact calculator and side-by-side presentation:

$$Y_{ip} = \beta_0 + \beta_1 T_{ip} \cdot I_{ip} + \beta_2 T_{ip} + \beta_3 I_{ip} + \delta_i + \alpha_p + \varepsilon_{ip}$$

where i indexes respondents evaluating program type p:

- Y_{ip} is our primary outcome of interest, i.e. the scaled log("valuation" or perceived program value) as defined above
- T_{ip} is an indicator equal to one if this program screen was a treatment screen
- I_{ip} is the scaled log(program impact), as defined above
- δ_i captures respondent fixed effects
- α_p captures program fixed effects

In this specification, our main test of interest is whether:

- $\beta_1 > 0$, or that the pooled treatment increases sensitivity to a change in impact

Finally, we will report the coefficient on I_{ip} as a proxy for the range of sensitivity respondents exhibit in response to our program descriptions. However, this will be included only as a descriptive statistic.

Secondary Analysis:

We will run additional analyses to look at:

1. The relative sensitivity to our three impact features
2. Heterogeneous treatment effects by our three impact features
3. The relationship between certainty and sensitivity to impact
4. The impact of our treatments on certainty
5. The relationship between incentivized and unincentivized responses
6. The relationship between respondent characteristics, notably experience with evidence and evaluation, and sensitivity as well as certainty
7. The relationship between time spent on the program valuations and sensitivity

Follow-Up Analyses:

NA

Inference Criteria, Including Any Adjustments for Multiple Comparisons:

We will use simulations to generate corrected *p*-values that control for the family-wise error rate across our two core primary tests, i.e. the test of the efficacy of the impact calculator and side-by-side presentation. We will not additionally correct for the inclusion of the pooled treatment indicator.

We will not include corrections for the secondary analysis, which we will clearly distinguish as secondary.

Limitations:

Relevant limitations are already identified in the design document.

Exploratory Analysis:

We will also look at how features of our analysis vary by agency. Does baseline sensitivity vary by agency? Do we see differences in experience with evidence and evaluation across agencies?

Similarly, we also plan to use the end-of-survey questions to understand the process of moving from evidence to action within each agency involved in this project. This will involve qualitative analysis of open-ended text responses.