Analysis Plan

Project Name: Incorporating evaluation into digital forms
Project Code: 2118
Date Finalized: July 6, 2022

Project Description

This project aims to build capacity to incorporate rigorous testing into digitization of federal forms. The goal of this initial proof-of-concept pilot is to test whether the way in which instructions are presented (i.e., as a text block at the start of the form versus embedded with the relevant question) affects how people complete the form. We will look for differences across these experimental conditions in the distribution of responses and form related outcomes, such as completion rate and time to complete form. A larger goal of this project is to set up a system for consistent randomization and data analysis in this online form setting that allows for conducting well-powered experiments.

The sample will be recruited from members of the general public using links and a short recruitment message sent via email newsletters and posted on assorted government websites that are commonly visited by members of the general public.

Once individuals click on the link to take the survey, they will be randomly assigned to one of two form conditions:

1. Instructions up-front: All instructions for completing the form will be provided on the first page of the form. (Form A)
2. Embedded instructions: All instructions for completing the form will be embedded with the relevant questions, and can be displayed by clicking a "+" next to the instruction box on each page. (Form B)

Other than the instructions format, all other questions and language on the two form versions will be identical.

Preregistration Details

This Analysis Plan will be posted on the OES website at oes.gsa.gov before outcome data are analyzed.

Hypotheses

We have no a priori hypotheses about the direction of the effect on any of the outcomes listed below.

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):

Data will be response data collected directly from users via the form, as well as metadata on each form (e.g., number of form visits, number of incomplete form submissions).

Outcomes to Be Analyzed:

Primary Outcome

1. Form submission: Of respondents who initiate the form, percent who submit it.

The following are secondary outcomes because their interpretation depends on the result of the primary outcome test. If there are differences in submission rates between conditions, due to differential attrition we will be unable to analyze secondary outcomes. If there are no significant differences in form submission rates, we will analyze secondary outcomes to determine whether the two different form types produced different types of responses.

Secondary Outcomes (which can be interpreted if submission rates do not differ)

1. Missingness: Percent of non-demographic questions skipped by respondents who submit the form. A skipped question will be defined as a missing answer (i.e., "?" or nonsensical answers will not be counted as skipped).
    a.
2. Form completion time

*Exploratory Outcomes*Distribution of reported income

3. Number of household members entered on part B of the form (income by household member).
4. Percent of "Unsure" responses to four questions about household members (has dependents, claimed as a dependent, share expenses, buy or prepare food together).

Tentatively plan to analyze this outcome, but uncertainty given data constraints and extra page for Form A

5. Page views (number of pages viewed, normalized by the number of pages in the form

version and adjusted for the number of form initiations) as a proxy for partially completed forms.

Imported Variables:

We anticipate merging three datasets:

- Responses and metadata (e.g. start date/ time, end date/time) for Form A
- Responses and metadata for Form B
- Data on how many users were sent to Form A and Form B (from Google Optimize) and on pageviews for each form (from Google Analytics)

Values from the third dataset will be used as the denominator to compute form completion, where the numerator is the number of lines (submissions) in each of the first two datasets. Note that depending on the data we are able to extract from Google Optimize, we may not have a precise value for the number of users sent to each form version. In this case we will assume a 50/50 assignment to form version and will use that number as the denominator to compute form completion.

The first two datasets will be combined so that we can compare responses between Form conditions.

Transformations of Variables:

- Treatment indicator: We will create a binary variable that indicates whether responses came from Form A or Form B.
- Completion time: If it is not provided in the form metadata, we will compute form completion time as the time difference between start time and end time, which are both fields in the metadata.
- Household income: Sum of all reported income values/sources for all entered household members
- Total number of household members: Number of household members for which information is entered in Part A
- Percent “unsure” responses: Total number of “Unsure” responses in Part A of the form divided by (count of Y responses + count of N responses + count of Unsure responses)
- Missingness: Total number of missing responses in Part A and Part B of the form (for all submitted forms) divided by (10 x the number of household members entered).

Transformations of Data Structure:

None

Data Exclusion:

We will also exclude responses who meet the following criteria:

- Entered "self" in Part A and reported an age under 18 years
- Entered "self" in Part A more than one time
- Entered "self" in Part B more than one time

The only form questions likely to create outlier values are income and form completion time. Since our analysis compares the distributions of incomes (rather than, e.g., the mean value in each form condition), we can include these responses without transformation or exclusion (while acknowledging that outlier values of income may not be accurately reported).

For form completion time, we will define outliers as any response with a completion time more than three standard deviations above the mean for all responses. For the purpose of evaluating the effect of treatment on completion time, we will recode outliers to the maximum completion time based on mean +/- three standard deviations.

Treatment of Missing Data:

We will only receive submitted form responses (i.e., those where a respondent has clicked "submit"). Our primary outcome is unaffected by missing data (form submission rates). If we find that completion rates differ across form types, it will not be possible to attribute differences in our secondary outcomes to treatment assignment; secondary outcomes will therefore not be analyzed.
If treatment assignment does not cause differential attrition, we will analyze the secondary outcomes using pairwise deletion, including only those responses where the key outcome is not missing. (For instance, when we compare the distribution of incomes reported on the two form versions, we will only include forms that reported income.) The treatment of missing data in computing the outcome "Percent Unsure" is described above in the section on Transformation of Variables.

Descriptive Statistics, Tables, & Graphs

Range, mean, standard deviation, median for the 7 outcomes (1 primary, 6 secondary) described above

Statistical Models & Hypothesis Tests

Since the randomization is being conducted via Google Optimize, prior to conducting analysis, we will check for balance of available metadata variables across conditions including:

- Number of people who started the form;
- Mobile vs. desktop users;

- Geographic location;
- Day and time the form submission was started; and
- Other metadata variables that may be available.

Statistical Models:

We will create a dataset where each row represents a user sent to a form. We will do this by merging the form response datasets described above (submissions of Form A, submissions of Form B) and adding additional rows based on the number of form initiations not submitted as captured in the data from Google Optimize. We will use this individual-level dataset to conduct difference-in-means tests using regression models (plus the Kolmogorov-Smirnov test of distributions specified for Secondary Outcome 3). In each case the significance test on the coefficient that represents form condition tests whether that outcome differs between conditions. The distribution of the dependent variable (e.g., linear, negative binomial, etc.) will be specified after examining the distribution of the data.

Primary outcome

1. Form submission: Linear regression model with outcome the % of forms initiated that were submitted, and an indicator variable for form condition (Form A versus Form B).

Secondary outcomes

1. Form completion time: outcome is the completion time by those who submitted the form
2. Missingness: outcome is the percent of non-demographic questions skipped by respondents who submit the form
3. Distribution of reported income: Kolmogorov-Smirnov test of distributions comparing total income reported on Form A and Form B
4. Number of household members: outcome is the number of household members entered on part B
5. Percent of “Unsure” responses: outcome is the percent of Unsure responses
6. Page views (i.e. last page viewed; IF available): t-test comparing number of the last page viewed on Form A and Form B, only for forms not submitted.

Confirmatory Analyses:

The six tests above are confirmatory in the sense that we predict a difference between conditions, but we have no prediction about the direction of the difference.

Exploratory Analysis:

In addition to the primary analyses outlined above, we plan to conduct the following exploratory analyses:

- Earned income and unearned income separately

Inference Criteria, Including Any Adjustments for Multiple Comparisons:

All tests are two-tailed. We will use $p < .05$ as the critical value for statistical significance. Because there is only one primary outcome, we will not correct for multiple hypothesis testing in the primary test.

The 6 secondary analyses here are part of the same family (they all inform the same policy decision of how to present form instructions). Thus, we will correct for multiple hypothesis tests using the Holm-Bonferroni method. As described in the OES guidelines for multiple hypothesis testing:

1. Run all $m=6$ hypothesis tests. Collect p-values for each test.
2. Order the p-values from smallest to largest, such that $p_1 < p_2 < p_3 < p_4 < p_5$.
3. Consider p_1 statistically significant at the per test significance level $0.05/(m-i+1)$ where a. m is the number of hypotheses b. i is the rank of the p-value
4. Stop procedure as soon as p is no longer statistically significant at corresponding per test significance level.

Limitations:

- OES has ***less certainty about the sample size*** than for a typical project, as well as ***uncertainty about the likely effect sizes***. This has made power analysis challenging.
- The form being tested is hypothetical. ***Respondents*** will be people who are motivated to spend time doing something to help the government, and ***may not be representative of the intended population*** (e.g., individuals completing forms to apply for government benefits). Using this sample may limit external validity. At the same time, since these respondents may be motivated to respond carefully, this probably means that any identified effect underestimates the true effect in the population of interest. After careful consideration of options, this is a limitation we are willing to accept at this stage given that this is a pilot project.
- Given the hypothetical nature of the form, ***we are unable to examine accuracy*** of responses directly, although that is an outcome of interest. Instead we are looking at proxies such as completion rates and distribution of income (the latter may be related to accuracy but we cannot say for sure).

Link to an Analysis Code/Script:

N/A