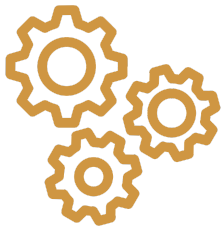
Analysis Plan

Project Name: Increasing naturalization rates for Lawful Permanent Residents

Project Code: 2121

Date Finalized: 2/14/2023



Project Description

Executive Order 14012, Restoring Faith in Our Legal Immigration Systems and Strengthening Integration and Inclusion Efforts for New Americans renews U.S. efforts to promote naturalization, raise awareness of the importance of citizenship, and advance and ensure equity throughout the citizenship and naturalization process. United States Citizenship and Immigration Services (USCIS) is committed to promoting naturalization of Lawful Permanent Residents (LPRs), a critical component for advancing democratic inclusion and increasing economic prosperity, through a variety of policy, program, and outreach efforts and building the evidence base on the effectiveness of its naturalization efforts. The proposed project will use a randomized control trial (RCT) to identify the effect of direct communication towards Lawful Permanent Residents (LPRs) on their likelihood of applying for naturalization (i.e., submitting the N-400 form). The primary approach will be to mail a letter to LPRs who are likely eligible for naturalization. This letter addresses various behavioral barriers by including: a fresh start motivator (framing moments in time as new beginnings), the benefits of US citizenship, a checklist of next steps, and a social motivator (information on peer behavior).

Preregistration Details

This Analysis Plan will be posted on the OES website at oes.gsa.gov before outcome data are analyzed. In addition, this project will be pre-registered in the American Economic Association RCT Registry at https://www.socialscienceregistry.org/.

Hypotheses

Sending letters that encourage eligible LPRs to naturalize will increase submissions of N-400 applications, the method by which individuals apply for naturalization. Secondary outcomes include the rate of online versus paper N-400 submissions and the usage of N-400 application fee waivers.

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):
USCIS will provide data on LPRs who are likely eligible to naturalize for citizenship and have not yet submitted an N-400. These data will include the following variables (all variables are from when an individual became LPR): a random identifier to anonymously track individuals over time, to ensure no individual can be identified; year of LPR status; class of admission (i.e., the process by which they were approved for LPR status); year of birth; country of birth; gender; marital status; geographic location (state, zip code, and CBSA); and USCIS estimated propensity to file an N-400. USCIS will also provide a separate dataset that identifies N-400 submissions, including date of submission and other ancillary details, which will serve as the outcome measure.

Outcomes to Be Analyzed:
The primary outcome is whether individuals submitted an N-400 application, which is the form that LPRs use to apply to naturalize as a U.S. citizen. Secondary outcomes include whether the N-400 submission is an online or paper form submission and whether the individual used an application fee waiver. Application fee waivers are only available for paper submissions.

Imported Variables:
Our primary approach is to block randomize by two variables that are provided in the data: (1) class of admission and (2) years since LPR status. This approach is to provide a basis for estimation of heterogeneous effects of the experiment by these two categories of individuals. Class of admission refers to the method by which an individual became an LPR and will be aggregated into five categories: immediate relative of U.S. citizen (class of admission labeled as Parent, Spouse, Children in data); family-sponsored preferences (Family); employment-based preferences (Employ); refugees and asylees (Humani), and; all other (Other). Our sample will consist of all individuals who became Lawful Permanent Residents from 6 to 9 years prior to the outreach, and we are interested in disaggregating treatment effects based on the specific year since designated LPR (i.e., there will be four groups: 6, 7, 8, 9 years since LPR). In addition, we will block randomize by an individual's gender and marital status (divided into three categories: Single, Married, or Other) at time of receiving LPR status. This approach is taken to ensure better balance across treatment and control samples; initial results suggest the smallest strata will have 400 individuals, though most are much larger.

Transformations of Variables:
There will be essentially no data transformation as most data used in the analysis will be provided in the final form. Class of admission will be slightly edited to be aggregated to the five main categories. Marital status will be combined into three groups of Married, Single, and Other. The primary transformation will be to classify the outcome variable of a N-400 submission to encompass the period covering six months from when the letter is sent to the treatment group. Given that not all letters will be sent simultaneously, this variable will take into account the date when an individual would have been sent a letter. Given USCIS processes, the letters will be sent on a rolling basis over roughly 6 to 8 weeks to the full sample of 300,000 treated individuals, so 6 months post randomization will vary by individuals.

Transformations of Data Structure:
There are no plans to transform the structure of the data. Primary outcome measures will need to be confined to the six months from the date the letter is sent to individuals.

Data Exclusion:
The data transferred from USCIS to GSA will consist of all LPRs that have never submitted an N–400. We will restrict to individuals who are 6 to 9 years since their LPR status was approved and will exclude any individuals whose years since LPR status date does not meet these criteria. Additionally, LPR subpopulations who require special protection such as T Visa, U Visa, and Violence Against Women Act (VAWA) recipients will not be included.

Treatment of Missing Data:
Individuals missing any of the following variables will be excluded from the randomization process: date of LPR status; class of admission. Covariates included in the analysis that are missing data will have a "missing" dummy included in the regression analysis.

Descriptive Statistics, Tables, & Graphs

Our primary tables will include:
1. A table showing balance among treatment and control individuals on the set of background variables made available, which may include: current age; country of birth (dummy variables based on the five largest groups to be determined, such as Mexico and China); and USCIS estimated propensity to file an N-400. Except for age and propensity to file, all variables are what was reported when the individual was originally granted LPR status.
2. A table showing estimated "intent to treat" effects of being sent a letter based on randomized treatment assignment. This table may also show IV results of treatment assignment based on a first-stage of whether the letter was not returned to USCIS as undeliverable (i.e., where the first-stage is having a valid address that did not reject the letter delivery), depending on whether that information is made available.
3. A table showing heterogeneous treatment effects based on (A) the five categories of class of admission and (B) years since LPR status was approved (6, 7, 8, or 9 years).

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:
Treatment assignment rates will be identical within each strata. We will estimate the rate at which we can assign treatment as the total number of letters divided by the total sample size (e.g., 300,000 letters to be sent over a sample size of 3,000,000 eligible LPRs would lead to a 10%

treatment assignment rate), then assign that percentage of individuals to treatment separately within each strata.

Our main specification is a linear model shown in equation 1 that includes an indicator for treatment assignment and indicators for each of the 120 strata ("years since LPR" by "class of admission" by "gender" by "marital status"; 4 * 5 * 2 * 3 = 120 categories).

$$y_{is} = \beta \cdot T_{is} + \mu_s + x_{is} + \varepsilon_{is} \quad (1)$$

In this specification we have an outcome y_{is} which indicates whether an N-400 was submitted for individual *i* in strata *s*. We then have T_{is} as a dummy which equals 1 if the individual was randomly assigned to receive a letter within that strata. Our primary controls are the strata fixed effects μ_s and then x_{is} which controls for age and country of birth. To the extent possible, we will bin and dummy these variables (e.g., age 20-24, 25-29, etc.) rather than use any continuous measures. We will use the Lin (2013) approach where we demean the covariates and interact the treatment with the demeaned covariates.

Our primary interest is in observing the coefficient β on our treatment dummy, which estimates the difference in N-400 submissions between our treatment and control groups. We will also estimate heterogeneous treatment effects by years since receiving LPR status and class of admission. In this approach, we will designate one of the strata as the reference category and then interact dummy variables for each of the remaining heterogeneous effects with the treatment assignment indicator. For example, we would run the following equation (2) to test heterogeneous treatment effects by LPR status:

$$y_{is} = \beta_1 \cdot T_{is} + \beta_2 \cdot T_{is} \cdot I(LPR = 7\,years) + \beta_3 \cdot T_{is} \cdot I(LPR = 8\,years) +$$

$$\beta_4 \cdot T_{is} \cdot I(LPR = 9\,years) + \mu_s + x_{is} + \varepsilon_{is} \quad (2)$$

In this case β_1 is our primary treatment effect for individuals who are six years since achieving LPR status, and the coefficients β_2, β_3, and β_4 estimate whether the treatment effect was significantly different for individuals who are seven, eight, or nine years since achieving LPR status. We will also run a similar regression based on class of admission, disaggregating by five categories. One note is that the smallest of the five categories of class of admission is estimated to be approximately 6% of the total data, and we will treat the point estimate on this last group as exploratory as we are underpowered to detect heterogeneous treatment effects for this group.

Confirmatory Analyses:
Our prior confirmatory analysis is to test the null hypothesis of whether our treatment effect is equal to zero at a standard significance level of 0.05. We will also perform similar tests for heterogeneous effects by years since LPR status and class of admission.

Exploratory Analysis:
At the request of USCIS, we will also estimate heterogeneous treatment effects by the additional covariates provided, including age, gender, propensity to file, marital status, and country of birth. We are likely to use causal forests and R-Learners to estimate general treatment effect heterogeneity (following the methods proposed in Athey, Tibshrani and Wager (Annals of Statistics, 2019) and Nie and Wager (2017)). We consider these investigations to be exploratory.

Inference Criteria, Including Any Adjustments for Multiple Comparisons:
We will be using a two-tailed, t-test of the estimated coefficient of the treatment assignment indicator at a standard significance level of 0.05. We will also be testing for heterogeneous effects for years since LPR status and class of admission. Based on our power calculations we do not plan to control for multiple hypothesis testing. There is no minimum effect of interest, as USCIS is interested in whether there is any evidence that the experiment increases N-400 submissions.

Limitations:
We do not anticipate any limitations in the proposed analysis. The strongest concern is whether the letters will reach their intended audience, as low rates of mail delivery will downwardly bias our estimates towards zero. We anticipate receiving indicators of mail being rejected, which will be used to estimate IV results based on having a valid physical address for mail delivery. Having a valid address does not necessarily indicate that the intended individual received or read the letter, but having an invalid address strongly guarantees that it was not received.

Link to an Analysis Code/Script:
We will provide the randomization code once the data are received and randomized. We will also generate a placebo outcome and estimate results based on the placebo outcomes, which will show our estimating equation.