# Analysis Plan

Project Name: Increasing FAFSA completion among HUD-assisted youth: Phase II

Project Code: 1717

Last Updated: 10/23/2018

---

## *Background*

Completing the Free Application for Federal Student Aid (FAFSA) is the first step to receiving financial aid, but many youth who qualify for financial aid do not receive it because they do not complete the FAFSA. The are several reasons why people do not fill out the FAFSA, including lacking general knowledge about the financial aid process, having difficulty with the complexity of the FAFSA, not having access to parental financial information, not having easy access to the FAFSA, and procrastination. These barriers can be particularly pronounced in low-income communities. Low-income students are more likely to be the first in their family to go to college, and their schools may have fewer financial aid and college-going resources, including college counselors, leaving students with a lack of information about the actions they must take. Low-income families may also lack a computer and reliable access to internet in their homes, which makes it more difficult to take advantage of the simplifying features of FAFSA on the web.

Part of the U.S. Department of Housing and Urban Development's (HUD) mission is to utilize housing as a platform for improving quality of life, including by increasing educational opportunity. HUD partnered with the Office of Evaluation Sciences (OES) in 2017 to field two tests of communications encouraging youth in public housing to complete the FAFSA.

The first test was fielded at the New York City Housing Authority (NYCHA). Public housing residents 17-24 years old were randomly selected into an intervention group which was sent a series of FAFSA-related communications or a control group that was not sent any information. The intervention group was sent a series of four communications via postal mail and a robocall. Households with an email address on file were also sent a series of three emails which mirrored the content of the first three mail-based communications.

Table 1: NYCHA Intervention Components

| Method (date sent) | Content |
|---|---|
| Letter in mail (February 15, 2017), email (March 2, 2017) | Message from NYCHA CEO with simple steps for starting the FAFSA online. |
| Robocall (March 3, 2017) | Complete the FAFSA to be eligible for federal student aid. |
| Letter in mail (March 7, 2017), email (March 9, 2017) | Pell Grant amount compared to CUNY tuition with simple steps for starting FAFSA online. |
| Postcard (March 29, 2017), email (April 11, 2017) | It's not too late message with implementation intention prompts. |
| Letter in mail (April 19, 2017) | Five FAFSA facts to counter common myths. |

The second test took place in spring 2017 at the Seattle Housing Authority (SHA) and the King County Housing Authority (KCHA) in Washington State. All residents in SHA and KCHA between 17 and 24 years old were mailed two letters. The letters were similar to the first two letters sent in NYCHA. The first letter included a statement from the CEO of SHA and KCHA, respectively, and two steps for creating an FSA ID and starting the FAFSA online. The second letter included a header with the maximum Pell Grant award and how a full Pell Grant compared to the average tuition at Washington schools. The second letter also contained the same two steps for creating and FSA ID and starting the FAFSA online.

## *Research Questions*

The primary research question is:

1. Can light-touch communications change FAFSA completion rates for youth living in Public Housing?

Secondary research questions are:

2. Can light-touch communications change post-secondary enrollment outcomes for youth living in Public Housing?
3. Are there subgroups (e.g., females or certain age ranges) which are more or less influenced by the communications?

## *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 Sources:**

*HUD Public and Indian Housing Information Center (PIC)*

HUD creates a data extract for research at the end of each quarter (March 31, June 30, September 30, and December 31). The data come from the HUD-50058 form that Public Housing Authorities complete for all residents upon initial admission, in the case of any moves, and when the household recertifies eligibility, which generally is an annual requirement. The files include data for each member of the household (see Section 3) and also household-level information.

*Department of Education Enterprise Data Warehouse and Analytics (EDW&A)*

The Department of Education (ED) collects data for all completed FAFSAs as well as loan disbursement information and post-secondary enrollment information. The FAFSA data include the academic cycle for which the FAFSA was completed, the date the FAFSA was completed, and the list of colleges to which to the applicant sent the FAFSA. Loan disbursement information includes the types of federal loans for which a student qualifies, and loan disbursement types, amounts, and dates. Enrollment status is reported by institutions and will include the institution name, type (public, private, proprietary), length of program (4-year or 2-year), and enrollment status (e.g., full time, half time, less than half time).

**Outcome Variables to Be Analyzed:**

FAFSA completion during the 2017/2018 award year is the primary outcome of interest. Secondary outcomes of interest include:

- FAFSA completion during the 2018/2019 award year,
- Pell Grant receipt during the 2017/2018 and 2018/2019 award years,
- auto zero Pell Grant receipt during the 2017/2018 and 2018/2019 award years,
- post-secondary enrollment during the 2017/2018 and 2018/2019 award years,
- institution type (private, public, or proprietary) during the 2017/2018 and 2018/2019 award years, and
- program length (2-year or 4-year) during the 2017/2018 and 2018/2019 award years.

**Transformations of Variables:**

The outcome data are owned by the Department of Education (ED). The terms of the data sharing agreement between the Department of Housing and Urban Development (HUD) and ED stipulate that HUD will send ED a pass through file. The pass through file will contain a list of individuals with names, dates of birth, and Social Security Numbers to facilitate the match. Additionally, the pass through file will contain variables to be used in the analysis, including demographic

information, variables used in the design of the randomization, and indicators of treatment status. ED will aggregate data as requested by OES to produce tables that can be used for simple analyses. The table specifications are detailed in the matching document. ED will check all aggregate tables before sharing them to make sure cell size counts are sufficient to prevent individual re-identification. If it becomes apparent that the cell size rules will be violated, the grouping rules will be amended.

### *NYCHA transformations:*

OES is requesting cell counts created by the intersection of several factors of substantive interest:

- Treatment indicator (0 for control; 1 for treatment)
- Email (0 none on file, 1 email address on file)
- Age (integers 17-24)
- Sex (0 for male, 1 for female)

There are 50,934 unique individuals between 17 and 24 years old in 39,484 unique households in the sample. While the large sample suggests cells will be fairly large, there may be fewer individuals in the older age groups. If cell size constraints require collapsing some grouping variables, the first step will be to collapse the age variable into two-year age intervals (i.e., 17-18, 19-20, 21-22, and 23-24).

Note that these tables can be used to produce point estimates, but they will understate the true standard errors and should not be used as a confirmatory analysis. The randomization was performed at a household level, and standard errors should be clustered at the household level. Analysis based on the aggregate tables will be used both as a validity check on the estimates produced by regression analysis and as exploratory analysis for subgroup effects.

### *Seattle and King County transformations:*

The unit for analysis for Seattle and King County will be at the level of the PHA. As such, tables only need to include aggregated counts for each PHA-year combination. Each PHA-year cell count should be based on only those residents who were living in the PHA for the relevant year and were between 17 and 24 years old. In general, residency will be considered as of March 31 of the relevant award year. which is the academic year that begins approximately 6 months later. For example, the outcomes used for residents residing in a PHA as of March 31, 2018 will pertain to the 2018/2019 award year. The exception to this rule is for the two treatment PHAs, whose participants were taken from the Dec 31, 2016 files.

HUD will have the relevant information to calculate the total number of eligible individuals in each PHA as a check on the aggregation performed by ED and will also be able to aggregate any relevant demographic characteristics included in the PIC to be used as PHA-level covariates.

**Imported Variables:**

### *Seattle and King County*

It will be necessary to join PHA-level aggregated counts of demographic variables from PIC back to the aggregated FAFSA completion data provided by ED.

**Transformations of Data Structure:**

See the Section: Transformation of Variables.

**Data Exclusion:**

None

**Treatment of Missing Data:**

The nature of the data match does not allow for accurate identification of missing data. ED collects the social security numbers, names, dates of birth, and other identifying information for anyone who completes the FAFSA. HUD will provide ED with the Social Security Numbers, names, and dates of birth for individuals in PIC who were included in the pilots. There are two potential sources of missing information or mismatches. One possibility is a partial match in which the social security number in one file matches a social security number in the second file, but other information such as name and date of birth do not match. While ED does not provide a detailed explanation of its matching process, we expect that given the verification processes involved in both systems the prevalence of mismatches will be small. There are no plans to account for any data mismatches in the analysis, but we will report, if possible, how prevalent the problem is. A second possibility is that a person who is living in one of the PHAs is not included in PIC. For example, the household includes a person who is not put on the lease for some reason. We are not able to observe these individuals.

## *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:**

### *NYCHA statistical models:*

We can reconstruct an individual level dataset with the cell counts and complete t-tests on the difference of means for the outcomes based on treatment status. Unfortunately, this model will not account for the clustering of standard errors at the household, which is an artifact of the randomization procedure.

While blocks were used in the original randomization (above or below the 95th percentile of household income, age, and development), given the equal probability of selection into treatment

for all units, the simple difference of means estimator is unbiased – meaning the point estimate will in expectation be correct – but because the standard errors will be too small, estimates from the aggregated cell counts will not be used to determine statistical significance but rather as a validity check on the regression output.

OES has requested that ED run several regressions on the individual-level data which take into account proper calculation of standard errors. Randomization was blocked on an indicator for being above or below the 95th percentile of household income, housing development, and whether or not the household had an email address on file. The first requested regression will be a simple regression of the outcome on an indicator for treatment ( $T_i$ ) and blocking variables ( $X$ ). The second will add additional covariates ( $Z$ ) to the first model: age, relationship, gender, race, ethnicity, household income, and total 17 to 24 year olds in the household. A third regression will add interaction effects between treatment and age, treatment and gender, and treatment and email to estimate subgroup effects.

(1) $$y_i = b_0 + b_1 T_i + \delta X + e_i$$

(2) $$y_i = b_0 + b_1 T_i + \delta X + \gamma Z + e_i$$

(3) $$y_i = b_0 + b_1 T_i + b_2 T_i \times Age_i + b_3 T_i \times Gender_i + b_4 T_i \times Email_i + \delta X + \gamma Z + e_i$$

### *Seattle and King County statistical models:*

A synthetic control method will be used to compare the estimates of FAFSA completion in SHA and KCHA to synthetic comparison groups constructed as weighted composites of other PHAs. The donor pool will consist of the 100 largest PHAs based on the number of public housing units available in the first quarter of 2016, excluding SHA, KCHA, and NYCHA. ED will provide PHA-level outcomes for each award year from 2008/2009 through 2018/2019. The post-secondary outcomes will be used in combination with other PHA-level characteristics to create synthetic comparison PHAs for SHA and KCHA. A different synthetic comparison will be created for each outcome.

We will use the R package Synth, used in Abadie, Diamond, and Hainmueller (2010), to create the synthetic control groups. The basic process in Synth uses an algorithm to optimally weight the donor PHAs based on prior year outcomes and other covariates to create a composite PHA that best mirrors the characteristics and outcomes of the treated PHA. The composite PHA is then used as a comparison to the treatment PHA for outcomes in the post-treatment period to estimate the effect. The estimator will be a simple difference in means comparison between the focal PHA and the relevant synthetic comparison PHA.

(4) $$ATE = \bar{Y}_{treat} - \bar{Y}_{synth}$$

P-values will be created from a permutation-style analysis with will create a null distribution by repeating the optimal weighting procedure for each PHA in the donor pool.

**Follow-Up Analyses:**

No follow up analyses are planned other than those stated above. If we decide to conduct any additional analyses they will be clearly described as exploratory.

**Inference Criteria, Including Any Adjustments for Multiple Comparisons:**

### *NYCHA inference criteria:*

We will use the standard decision rule based on alpha = 0.05. We will present the effects on FAFSA completion as the primary outcome. Our preferred model will regress FAFSA completion in the 2017/2018 award year on an indicator for treatment and a set of blocking variables (equation 1). Regressions with additional covariates will serve as a robustness check on the main specification. The analysis based on aggregate cell counts will serve as a robustness check and may allow for exploratory subgroup analyses.

Secondary analysis will include the additional outcome variables and all outcomes in the 2018/2019 award year. We also plan to complete exploratory subgroup analyses based on age, gender, and the availability of an email address (which indicates a higher treatment "dose"). We do not plan to make any multiple comparisons corrections for the secondary analyses and will treat them as exploratory.

### *Seattle and King County inference criteria*

A permutation-based approach will be used to create p-values, and we will use the standard decision rule based on alpha = 0.05. The primary outcome of interest will be FAFSA completion in the 2017/2018 award year. Additional outcome variables will be treated as exploratory analyses. We do not plan to make any multiple comparisons corrections to the secondary analyses.

**Limitations:**

We will not be able to observe outcomes for people who do not match in the data. If for some reason an individual in the HUD file does not exactly match to an individual in the ED file, we cannot be sure if it is due to a data error (e.g., a typo in the SSN or mispelling of a name) or if it is a true non-match which would indicate the person has not completed the FAFSA. However, given that both ED and HUD attempt to verify all forms of identification as part of the eligibility process, we expect the number of non-matches to be relatively small.

The primary limitation of a synthetic comparison model is that it cannot account for any contemporaneous changes in the treated PHAs. It may be possible that some other policy change in the Seattle area went into effect during the same time as the communications campaign. Given that Seattle is contained within King County, this concern is not fully mitigated by estimating the effect for each PHA, separately.

There is an unlikely potential for movement of students more likely to complete the FAFSA into Seattle or King County (or of those less likely to complete the FAFSA out of the area) concurrent

with the evaluation; however, we have no reason to expect large population shifts and turnover in public housing has historically been low.

**Exploratory Analysis:**

No other exploratory analyses are planned in addition to those stated above. If we decide to conduct any additional analyses they will be clearly described as exploratory.