Analysis Plan

Project Name: Using Incentives to Reduce Nonresponse Bias in the American Housing Survey (AHS)
Project Code: 1901
Date Finalized: January 15th, 2022

Contents

1 Project Description

The American Housing Survey (AHS) is a biannual, longitudinal survey of housing units designed by the U.S. Department of Housing and Urban Development and administered by the U.S. Census Bureau. The sample of housing units is drawn from residential units in the United States and is designed to provide statistics that represent both the country as a whole and its largest metropolitan areas.

As with many federal surveys, the AHS has experienced declining response rates, requiring increasing amounts of time and effort to reach the 80 percent response rate preferred by the Office of Management and Budget. In particular, response rates have declined from approximately 85 percent in the 2015 wave to 80.4 percent in the 2017 wave to 73.3 percent in the 2019 wave.

As response rates decline, issues pertaining to data quality become increasingly important. While not indicative of bias in itself, a lower response rate can raise concerns that there is a correlation between the likelihood of nonresponse and survey items of interest. Nonresponse bias not only can diminish data quality by providing an inaccurate picture of the world, but also can diminish data quality by creating an over-reliance on post-survey adjustment procedures that add to the noise around population estimates even when recovering population estimates that are accurate. This project seeks to experimentally test the use of targeted monetary incentives to improve the quality of AHS data and to learn which methods of allocating incentives are most cost effective at increasing data quality.

When referring to nonresponse bias, we mean a divergence between a population quantity of key interest—such as the true proportion of U.S. adults living in severely inadequate housing—and its sample estimate, which arises due to systematic differences between those who do and do not respond to a survey. In theory, it is possible to adjust survey estimates to account for differential nonresponse so that sample estimates converge to population quantities, and bias is removed. To account for potential nonresponse bias, the AHS calculates a noninterview adjustment factor (NAF) that reweights for nonresponse within cells defined by metropolitan area, type of housing unit, block group median income, and area-level rural/urban status. In principle, adjustments such as this, along with raking,[1] should reduce or even remove the inferential threats posed by nonresponse bias. However, there is no guarantee that the model used for bias-adjusted estimates contains all the information it needs. Moreover, the weights used in such bias adjustment schemes typically increase variance in estimates: they essentially require units in grid cells with a lot of missingness to "represent" more unobserved units than those in grid cells with less missingness.

Furthermore, our preliminary analyses leave open the possibility that the raking and nonresponse adjustment factors currently employed to reweight AHS estimates do not ensure convergence with population quantities. In a separate memo on nonresponse bias in prior rounds of the AHS (see attached), we found two sets of systematic differences—nonrandom attrition from the panel; differences between sample quantities and known population quantities—that persist in spite of weighting meant to account for nonresponse bias. For the first, as an example, a key outcome the AHS measures is housing inadequacy. Among units where an interview was successfully conducted dur-

1. The AHS raking procedure, as implemented in the 2019 wave, is described in Section 3.4 of (U.S. Census Bureau and Department of Housing and Urban Development 2020). Broadly, this involves using "control totals"—or known estimates of housing and population totals from other sources—to adjust the weights on AHS respondents so that the AHS sample estimate of the housing or population characteristic moves closer to the control/independent estimate. Since moving sample estimates closer to control/independent estimates on one attribute (e.g., number of vacant housing units in a state) can mean sample estimates move *further* from population estimates for other attributes (e.g., number of persons aged 65+ in a state), the AHS defines a priority order for adjustment.

ing the 2015 wave of the AHS, some dropped out due to nonresponse in 2017. Reweighted estimates suggest 12 percent of those who stayed in the panel in 2015 and 2017 had problems with rodents. Looking at those housing units that appeared in 2015 only to drop out in 2017, however, only 9 percent had problems with rodents—in other words, a key measure of housing quality appears correlated with differential panel attrition. For the second, we found the AHS bias-adjusted estimate of the proportion of householders in the U.S. who own their home outright (without a mortgage or loan) in 2015 is seven percentage points lower than the corresponding proportion in the 2010 Decennial census count. Attributing these divergences to nonresponse bias with complete certainty is a challenging task since, by definition, we cannot measure the attributes of units who do not respond. However, the evidence presented in the nonresponse bias memo suggests that, in addition to adjusting sample estimates on the backend, improving sample composition on the frontend would increase the accuracy of estimates.

The purpose of this project is to determine whether and how the provision of cash incentives prior to contact with Census Bureau staff can reduce nonresponse bias in (adjusted and unadjusted) sample estimates. Furthermore, this test of incentives is intended to generate actionable evidence on the optimal way to target incentives—both how much and to whom—so as to maximize data quality and cost effectiveness.

Our intervention consists of sending cash to potential respondents sampled as part of the Integrated National Sample of the 2021 American Housing Survey. The cash was delivered inside an envelope containing a letter reminding the potential respondent about the survey. This letter was sent both to treatment and to control respondents, albeit with a slight wording change that mentions the incentive in the treatment letter and not in the control.

While many studies of incentives randomize differing amounts as does ours, a key innovation of this study is to randomize the method through which incentives were allocated. In one randomly-selected half of the sample, incentives were provided completely at random. In the other half, incentives were deterministically provided to the respondents estimated to have the highest likelihood of not responding. The method for estimating propensity to respond is described in greater detail in Appendix Section 5.1.

Because the very method used to allocate incentives is randomized, we can estimate the causal effect of using a propensity-determined versus a propensity-independent allocation method. In this document, we refer to the variable that assigns respondents to either of the two incentive allocation methods as T, for "targeting." When $T = 1$, the potential respondent receives the incentive allocation they would receive if propensity-determined allocation were used for the whole sample, and when $T = 0$, the potential respondent receives the incentive allocation they would receive if incentives were assigned completely at random.

Conditional on being allocated any incentive, potential respondents are randomly assigned to an amount of 2, 5, or 10 dollars. We denote this variable in this document using A, for amount. Appendix Section 5.2 describes the randomization procedure and justification for the incentive amounts. Table 1 describes the sample size in each condition, with $N = 86,017$ in the overall sample.

Table 1: Sample sizes per condition

Random assignment of **T** (incentive allocation method)							
Propensity-Independent (50%)				Propensity-Determined (50%)			
N = 43,009 (50%)				N = 43,008 (50%)			
Random assignment of **A** (dollar amount received)							
$0	$2	$5	$10	$0	$2	$5	$10
30,107 70%	3,225 7.5%	3,225 7.5%	6,452 15%	30,107 70%	3,225 7.5%	3,225 7.5%	6,451 15%

2 Preregistration details

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

3 Hypotheses

The research design is tailored to address a family of questions on how different kinds of incentive schemes affect nonresponse bias (a measure of data quality) and the effort to achieve that reduction in bias. Here, we outline the general sets of hypotheses and in Section 4.8 we discuss the estimands for each hypothesis and estimation strategy in greater detail.

3.1 Primary hypothesis: impact on nonresponse bias

Defining nonresponse bias as the expected difference between non-adjusted AHS sample estimates and their corresponding population statistics, we ask:

- To what degree does allocating the entire incentive budget to respondents deemed at highest risk of nonresponse reduce nonresponse bias, as compared to a purely random allocation of incentives?

We hypothesize that the allocation of incentives to those deemed most at risk of nonresponse will reduce nonresponse bias.

3.2 Secondary hypothesis: measures of effort to achieve data quality

While the main focus of the experiment is improving the quality of sample data, a secondary question of interest—holding data quality constant—is to understand the extent to which an incentive changes the level of effort required to achieve that data quality. We investigate the following question about the experiment's impact on the degree of effort it takes the survey to achieve high sample quality:

- What is the relationship between the amount of the incentive provided and the probability of nonresponse and number of contact attempts? Are there diminishing returns to the effectiveness of incentive amounts?

We hypothesize that targeting incentives to those at risk of nonresponse may not only lead to higher quality data (reduce nonresponse bias) but also may decrease the effort required to obtain that data. Focusing on incentive magnitudes, we further hypothesize that incentive amounts exert a monotonic positive effect on the probability of response and a monotonic negative effect on the number of attempts and time spent on a case. We expect that there may be diminishing returns to larger incentive amounts.

4 Data and data structure

4.1 AHS internal use files

Throughout the analyses, we primarily use the AHS internal use files (IUF) that contain information about both responders and nonresponders. We focus on the AHS national sample, a nationally-representative biannual panel. The AHS national sample can be classified into four exclusive categories: regular occupied interviews, in which the usual occupants of a unit are interviewed; a vacant interview, in which the owner, manager, janitor, or knowledgeable neighbor (if need be) of an empty building is interviewed; a "usual residence elsewhere" (URE) interview, for units whose occupants all usually reside elsewhere; and a noninterview. In the majority of analyses, we focus on contrasts between noninterviews (nonresponse) and the other three interview types (response).

Here, we review the main data sources. Unless otherwise specified, we use data from the 2021 AHS:

1. **Main IUF file for completed interviews:** for each respondent, this includes values for key attributes measured in the AHS (e.g., housing quality; demographic characteristics of the householder) as well as `J` flags for whether a particular variable has been imputed. This is used for the primary version of analysis one in Section 4.8, which focuses on the effect of the allocation method (propensity-determined versus independent) on nonresponse bias.

2. **Sampling frame and bridge files:** while the first data only contains values for respondents, these files, which vary across waves, contain sampling frame attributes known from addresses such as county-level rurality and housing type. We plan to use these data for (1) an alternate version of analysis one that examines characteristics measured in both respondents and nonrespondents and (2) the exploratory analysis of the impact on variance.

3. **Contact history file (CHI):** for each wave, we not only have a unit's response status but also have metadata on the field responders' attempts to locate and interview the unit. These fields include the number of times a unit was contacted (which can be several even among those who eventually respond), the dates between contact attempts, and other measures of the effort that went into trying to convert nonresponders into responders. We use these data for analysis two, which measures the impact of the propensity-determined allocation on measures of effort in addition to Yes or No response status.

4.2 AHS survey design and weighting

For the AHS national sample, the AHS uses a four-stage weighting procedure to generalize from the sample to the target population.[2]

In turn, there are three options for how we can use the weights from different stages of the adjustment process. These options correspond to two distinct quantities we report for different analyses: point estimates and measures of variance. They also reflect the fact that there are two sources of

2. First, analysts calculate a "base weight" (`BASEWGT`) that adjusts for the inverse probability that a unit is selected into the sample. Second, analysts apply so-called "first stage factors" (FSFs) that calibrate the number of units selected in each primary sampling unit strata to the number of housing units in these strata as measured using an independent Census Bureau estimate. The third stage involves a "noninterview adjustment factor"(NAF) that uses five variables to define cells for noninterview adjustment: Census division; type of housing unit; type of CBSA; block group median income quartiles; and urban rural status. The final step is applying what are called "ratio adjustment factors" (RAFs) to the weights through raking, which is designed to produce weights that lead to estimates with lower variance by calibrating weighted outputs to "known estimates of housing units and population from other data sources believed to be of superior quality of accuracy"(U.S. Census Bureau and Department of Housing and Urban Development 2018, 8).

variability in estimates from our experiment: (1) variability from the AHS sampling procedure and (2) variability from the experimental procedure. The options are:

1. **Report estimates without any weighting**: this would correspond to *point estimates* that represent sample rather than population quantities, since they do not account for the base weights and first stage factors (FSFs) that adjust for the sampling process. These weights are important for generalizing estimates to the survey's target population: a representative sample of the universe of U.S. residential housing units.[3] For this reason, all point estimates will be weighted (options two and three).

2. **Report estimates weighted by FSFs but variance estimates that only reflect variability from the experimental procedure:** in these results, the point estimates correspond to population point estimates but the variance on those point estimates only accounts for variability from the experimental procedure rather than variability from the AHS sampling procedure. Our main inferences will be based on this measure of variance.

3. **Report estimates weighted by FSFs and variance estimates reflect both variability from the experimental procedure and variability from the AHS sampling procedure:** we discuss this analysis in Section 4.11. This variance estimation involves using the FSFs and the 160 replicate weights corresponding to that weight.

4.3 Imported variables

In exploratory analyses, we may use 2020 tract-level American Community Survey (ACS) 5-year estimates to estimate the impact of propensity-determined allocation on contextual attributes.[4] Otherwise, none of the data is imported from external sources.

4.4 Transformations of Variables and Data Structure

We describe specific variable transformations when outlining how we define each outcome variable in Section 4.8. We do not anticipate changes to the data structure beyond the aggregations we performed for predictive modeling that we discuss in Section 5.1.

4.5 Data Exclusion

We do not anticipate excluding any data.

4.6 Treatment of Missing Data

We do not anticipate any missing data for the following outcomes: (1) sampling frame variables, (2) nonresponse (Y/N), and (3) number of contact attempts. For observed attributes among respondents (e.g., homeownership), we will treat missing as a distinct level for categorical variables and for continuous variables, will conduct mean imputation.

4.7 Statistical Models & Hypothesis Tests

4.7.1 Treatment conditions and probability weights

As described in Appendix Section 5.2, there are three variables that are randomly assigned: $T_i \in \{0, 1\}$ is an indicator for whether the unit receives the allocation they would have received under

3. More specifically: "The universe of interest for the AHS consists of the residential housing units in the United States that exist at the time the survey is conducted. The universe includes both occupied and vacant units but excludes group quarters, businesses, hotels, and motels. Geographically, the survey covers the 50 states and the District of Columbia (D.C.)"(p. 3 U.S. Census Bureau and Department of Housing and Urban Development 2020).

4. These will be available in March 2022.

the Propensity-Determined (versus Propensity-Independent) method; $Z_i \in \{0, 1\}$ is an indicator for whether the individual is assigned to receive any incentive amount in the allocation used; $A_i \in \{0, 2, 5, 10\}$ is the dollar amount allocated to each potential respondent.

The assignment procedure generates a correlation between the predicted probability of nonresponse and the probability of receiving an incentive. This correlation could cause bias if not accounted for, as it will result in the overrepresentation of certain covariate profiles and types of respondents in the incentive group. To correct for this issue, we weight units by the inverse of their propensity to be sent an incentive of any kind in any analyses that involve assessing relationships between incentive receipt and outcomes.

Specifically, since T is independent, for any given individual the probability of assignment is given by:

$$Pr(Z_i = 1) = Pr(T_i = 1)Pr(Z_i = 1 \mid T_i = 1) + Pr(T_i = 0)Pr(Z_i = 1 \mid T_i = 0).$$

For the 30% (m/n) of units with the lowest propensity to respond,[5] (who are allocated an incentive under targeting), this evaluates to $0.5 \times 1 + 0.5 \times 0.3 = 0.65$. For the 70% of units with the highest propensity to respond (who are not allocated an incentive under targeting), this evaluates to $0.5 \times 0 + 0.5 \times 0.3 = 0.15$. Thus, there are four possible values of a treatment assignment probability $\pi^Z_{i,z}$ (where z indicates a treatment status for respondent i):

1. For j low propensity to respond (high propensity to nonrespond) individuals:
    - Assigned to treatment (any incentive): $\pi^Z_{j,1}$ = 0.65
    - Assigned to control (no incentive): $\pi^Z_{j,0}$ = 1-0.65 = 0.35
2. For k high propensity to respond (low propensity to nonrespond) individuals:
    - Assigned to treatment (any incentive): $\pi^Z_{k,1}$ = 0.15
    - Assigned to control (no incentive): $\pi^Z_{k,0}$ = 1 - 0.15 = 0.85

As a result, it is possible to observe every unit in every treatment condition, albeit with differing probabilities. To obtain unbiased estimates of the average treatment effect of receiving incentives, we downweight those who are overrepresented in incentive or no-incentive groups, and upweight those who are underrepresented, using $\frac{1}{\pi^Z_{i,z}}$, the inverse propensity weight (IPW).

4.8 Confirmatory Analyses and Statistical Models

We plan to conduct three confirmatory analyses.

4.8.1 Analysis one: effect of propensity-determined allocation on nonresponse bias

This analysis focuses on key attributes of housing units, households, and areas measured by the AHS in 2021. This list, developed based on the nonresponse bias analysis we attach in the appendix and conversations with Census, will include the following variables from the IUF or sampling frame:[6]

1. Own house (no; yes with mortgage/loan; yes with no mortage/loan)

5. Or conversely, the highest propensity to not respond.

6. These variables derive from three sources. First are variables where 2015 AHS estimates deviated significantly from 2010 Decennial Census estimates (Figure 1 in nonresponse bias summary memo). Second are variables that are predictive of panel attrition using a large penalty term from a `LASSO` model (Figure 15 in the nonresponse bias summary memo). Third are sampling frame variables used in the nonresponse adjustment process.

2. Average household size
3. White alone (householder)
4. Age of householder
5. Rodents
6. Mold
7. *Sampling frame:* Census Division
8. *Sampling frame:* HUD-assisted unit (as of 2013)
9. *Sampling frame:* 2013 Metropolitan Area (county-level; principal city, nonprincipal city, micropolitan area, non-CBSA area)
10. *Sampling frame:* type of housing unit (house/apt; mobile home; other)

Conducting individual tests for each outcome would pose a multiple comparisons problem. Therefore, we conduct an omnibus test of the null hypothesis that the (conditional) difference in means between the propensity-determined and propensity-independent samples is zero across all outcomes.

Specifically, we conduct an F-test comparing a model in which the allocation method indicator, T, is regressed on pair-level block indicators and one in which T is regressed on pair-level block indicators and the list of outcomes above.

The F-test can be interpreted as a test of the null hypothesis that the true coefficients on the outcomes are all equal to zero. Rejection of the null hypothesis therefore implies that at least one of the outcomes is imbalanced with respect to T. Thus, we are able to run one test to understand whether the first moments of the distributions of any of the outcomes are different between the two different allocation methods.

Additional notes on estimation and inference are:

- We will conduct two versions of this analysis
  - **Main analysis:** this analysis is restricted to outcome variables from the list above that we only observe among respondents. Therefore, for this analysis, we subset to the respondent sample. The comparison is then between values for respondents under $T = 1$ and values for respondents under $T = 0$.
  - **Secondary analysis:** this analysis is restricted to outcome variables from the list above that we observe among both respondents and nonrespondents, since they represent sampling frame variables known prior to response. If the treatment changes values for these variables, it potentially reduced nonresponse bias.
- We will report the outcome-specific differences in means graphically but will not conduct inference on individual outcomes
- Similarly, due to different definitions across surveys and delays in the 2020 Decennial census, we will not try to systematically determine which group's values for a variable are more similar to those from a benchmark/target population. For instance, if our comparison of means shows that 60% of the control group owns their homes, while 63% of the treatment group owns their homes, we may contextualize these differences with reference to the national homeownership

rate measured in the Census ($\sim$ 65%). But our tests assess the between-group differences and not which group is closer to some external, benchmark value.

- We do not reweight using the IPWs discussed in Section 4.7.1 since T is independent of units' covariates and potential outcomes.
- We chose the F-test for two reasons. First, while the F-test only captures differences in the sample means (the first moment; e.g., the % of household heads who are White alone in the treatment and control groups), and not differences in quantities like the variance, our main focus is on differences in the sample means. This stems in part from the fact that most of the above variables we will include in the F-test are binary indicators (e.g., White alone or not; Mold or not) where the proportions reflect both the mean and the variance.[7] Second, while some raise concerns about the asymptotic properties of likelihood ratio tests in small samples (Hansen and Bowers 2008), our sample size is large enough ($\sim$ 84,000) for these properties to reasonably hold.

4.8.2 Analysis two: effect of propensity-determined allocation on response rate and effort

This analysis focuses on whether propensity-determined allocation improves two outcomes:

1. **Response rate:** we define this outcome as a binary variable where either a unit is an occupied interview (responder) with sufficient completeness to remain in the final IUF data file or not.
2. **Contact attempts:** we define this outcome as a continuous variable based on the CHI data and aggregating contact attempts across all modes.[8]

To estimate:

- We regress each outcome on pair-level block indicators and T
- Similar to the first analysis, we do not use IPWs since T is independent of units' covariates and potential outcomes
- **For inference:** we conduct randomization inference with $m = 5,000$ replicates and use a two-tailed p value.

4.8.3 How we will judge different patterns of results for analysis one and analysis two

Analysis one measures the impact of T on data quality. Analysis two measures the impact on the effort required to collect that data. We will interpret the combined results as follows:

1. $T = 1$ **increases response rate and leads to different sample composition:** the treatment increased response rate and may have reduced nonresponse bias.
2. $T = 1$ **increases response rate but does not lead to different sample composition:** the treatment increased response rate but had no detectable impact on nonresponse bias.
3. $T = 1$ **does not increase response rate but does lead to different sample composition:** the treatment changed the sample composition despite not changing the response rate.

7. There are some continuous attributes like age of householder, for which we might care about differences in the distribution of values even if there are no treatment-control differences in the mean. However, these are the minority of the list.

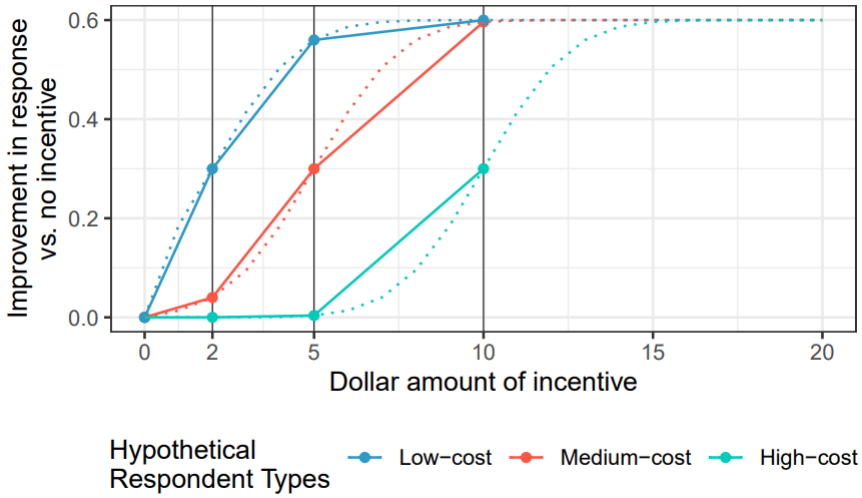
8. Another measure of effort is in-person contact attempts. We focus on all modes since it reflects phone-based effort as well.

4. $T = 1$ **does not increase response rate and does not lead to different sample composition:** the treatment neither increased the response rate nor improved nonresponse bias.

4.8.4 Analysis three: diminishing returns

This analysis will test for the presence of an inflection point in the relationship between dollars of incentives provided and probability of response and number of contact attempts. In particular, we are interested in whether incentives exhibit diminishing returns. Figure 1 illustrates hypothetical relationships between dollar amounts and response probabilities (dotted lines), alongside the linear relationships that will be estimable from the data, given the allocation of four incentive amounts ($0, $2, $5, and $10).

Figure 1: Example dose-response curves for different subsets of respondents



To do, we use the following estimation procedure, repeated across two outcomes (Y/N response; # of contact attempts):

1. **Test for diminishing returns from $0 to $2 versus $2 to$5:** we use a linear hypothesis test where the left hand side (LHS) represents the effect of increasing incentives from $0 to $2 and the RHS represents the effect of increasing the incentives from $2 to $5: $3 \times 2 - 3 \times 0 = 5 \times 2 - 2 \times 2$

2. **Test for diminishing returns as we increase incentive from $2 to $5 and $5 to $10:** we use a linear hypothesis test where the left hand side (LHS) represents the effect of increasing incentives from $2 to $5 and the RHS represents the effect of increasing the incentives from $5 to $10: $5 \times 5 - 5 \times 2 = 3 \times 10 - 3 \times 5$

Since these regressions involve A (randomized incentive amount) rather than T, we will employ the inverse of the probability weights described in Section 4.7.1, because receiving any incentive amount is correlated with units' potential outcomes.

To estimate:

1. We will use `lh_robust` within estimatr and car to specify the linear hypothesis tests

2. We will judge inference as $p < 0.05$

3. We omit one other other potential comparison—$0 to $2 versus $5 to $10—because (1) we assume the relationship is monotonic (if nonlinear), such that the response from $5 to $10 $\geq$ $5 to $2 $\geq$ $2 to $0 and (2) to reduce the total number of tests.

4.9 Exploratory Analyses and Statistical Models

4.9.1 Impact on effective sample size

As discussed earlier, there are two ways to address biases that arise from nonresponse:

1. Interventions to increase response rates/reduce nonresponse bias prior to data processing
2. Conditional on a given response rate, and during data processing, weighting to adjust for bias

One advantage of the first approach over the second is that weights, depending on their magnitude and distribution across units, increase the variance of sample estimates. We can summarize this issue using the concept of the AHS' *design effect*, or understanding how departures from simple random sampling and a perfect response rate affect sampling error in estimates.

To examine the impact of the experiment on variance in estimates, we take the following approach:

1. **Split the data into treatment and control and construct separate nonresponse adjustment weights:** split the data into two groups—$T = 1$ and $T = 0$—and loosely replicate the process that AHS survey designers use to create weights that adjust for nonresponse. This process is described in greater detail in Appendix 5.3.

2. **Obtain a point estimate for impact of those weights on effective sample size:** while one approach to comparing the weights is to examine how they influence variance around a particular statistic, another approach is to compare how they influence the effective sample size in each group. For this, we use Kish's approximate formula for computing effective sample size, calculating this value separately by group, where i indexes a respondent and w represents that respondent's weight created in step 1:

$$n_{\text{eff}} = \frac{(\sum_i^N w_i)^2}{\sum_i^N w_i^2}$$

3. **Find the difference in effective sample size between treatment and control:** we want the effective sample size to be as close to the nominal sample size as possible, so n_{eff} to be larger. We can calculate the following difference, and hope to see a positive value if the treatment improves the effective sample size, where $n_{\text{eff},1}$ represents the treatment group randomized to propensity-determined incentives and $n_{\text{eff},0}$ represents the control group:

$$\text{Diff sizes} = n_{\text{eff},1} - n_{\text{eff},0}$$

4. **Use randomization inference to judge statistical significance:** the previous step results in a point estimate of the difference in size. To judge whether this is statistically significant, we will repeat steps 1 through 3 $m = 5,000$ times permuting the treatment status to form a null distribution of differences in effective sample sizes. The p value will be a two-tailed test that measures (1) finds the percentage of permuted test statistics $\geq$ the observed test statistic; (2) finds the percentage of permuted test statistics $\leq$ the observed test statistic; and (3) takes the min of 1 and 2.

4.9.2 Heterogeneous effects of treatment

Given a finite budget of incentives, a key goal is to target those incentives to those for whom the incentive has the largest impact on whether they respond. For this, we shift from an estimand of the average treatment effect of incentives to the conditional average treatment effect (CATE) for each unit, or how the effect varies among units with different *pre-treatment* attributes.

For this analysis, we:

- **Focus on A (randomized incentives) and the contrast between any incentive and no incentive:** our reason is that it makes more sense conceptually to think of units that have different degrees of responsiveness to monetary incentives, rather than units with different degrees of responsiveness to the propensity-determined versus independent incentives. We collapse the different incentive amounts for statistical power reasons and because we believe the meaningful distinction is between some and none.
- **Restrict to respondents randomized to the propensity-independent condition:** while this restriction reduces the sample size, estimating CATEs among the propensity-determined condition, even with reweighting by IPWs, risks results that (1) find significant "moderators" of the treatment but that do so because (2) they were inputs to the nonresponse propensity scores discussed in Section 5.1. To reduce this possibility, we restrict to respondents randomized to the propensity-independent incentive condition.
- **Use machine learning (ML) methods to estimate CATEs using a high-dimensional set of pre-treatment attributes:** one approach to examining heterogeneous treatment effects is to use theory to select specific attributes that moderate the effect: for instance, a respondent's household income could be correlated with whether financial incentives shifts their response. Since we do not have strong *a priori* theory about what may moderate the effect, we will use machine learning to estimate the CATE. The pre-treatment attributes we will use will be similar to those used in the propensity score estimation (Appendix Section 5.1), including sampling frame variables, lagged nonresponse status, and categorical variables with nonrespondents set to a category of missing.[9] We are not prespecifying the ML estimation method we will use since certain methods may be more feasible than others within our computing environment, but options include `causal forest` (Wager and Athey 2018) or metalearners for CATE implemented in `causalToolbox` (Künzel, Sekhon, et al. 2019; Künzel, Walter, et al. 2019).[10]

4.10 Inference Criteria, Including Any Adjustments for Multiple Comparisons

In the sections above, we specified the inference procedure for each analysis. We will use $p < 0.05$ as the threshold for statistical significance. We do not plan to adjust for multiple comparisons because the number of tests remains small:

1. Analysis one: one omnibust test
2. Analysis two: two coefficients on T (one for response; another for contact attempts)
3. Analysis three: four linear hypothesis tests (two outcomes $\times$ two shifts in incentives)

9. We may also use ACS contextual data.

10. As described in (Künzel, Sekhon, et al. 2019), metalearners for the CATE involve two steps. First, the data is split into treatment and control: in our case, any versus no incentive within the $T = 0$ propensity-independent condition. Then, a "base learner"–or standard binary classifier—is used to predict the conditional expectation of the outcomes in each group. Finally, the algorithm finds the difference between the estimates in the treatment group and the estimates in the control group.

4.11 Robustness Checks

We plan to conduct two robustness checks.

First is re-estimating analysis two with a cutoff date to account for the Census stopping rule. The Census Bureau typically stops data collection once the target of an 80% response rate has been met. If incentives had been targeted at areas rather than specific respondents, this would pose risks of spillover effects—if we increase the response rate in area 1, then we may also decrease it in area 2 by reducing the need to collect more data there in order to achieve an 80% response rate. While our main way of addressing this is that we targeted incentives at respondents rather than areas, we will also work with Census to design a robustness check that:

1. Selects a stop date for when they devoted less effort to data collection due to response rates approaching 80% (if relevant)
2. Re-estimates effects as of or before the stop date

Second is to analyze the robustness of shifting from SATE to PATE. For the reasons discussed in Section 4.2, it is important to check that results are robust to the larger variances from incorporating the AHS sample selection and replicate weights. The main reason is that (1) the experiment may have heterogeneous effects and (2) there may be overlap between the pre-treatment attributes used in the AHS sample selection process (e.g., HUD-assisted as of 2013 or not) and pre-treatment attributes that effects are heterogeneous over. Therefore, examining robustness to the PATE may be important.

This entails using the FSF weights discussed in Section 4.1, which adjusts for the sample selection process but not for nonresponse bias, as well as the 160 replicate weights that correspond to that variable. For main analyses 1 and 2, which do not require inverse-propensity weighting, the procedure is relatively straightforward: we estimate the main results, weighting by the FSF. We then proceed through the 160 replicate weight vectors. At each vector, we rerandomize the treatment 1000 times, and employ the replicate weight as a weight in the regression. This provides 160,000 estimates of the PATE where the sharp null hypothesis of no effect for any unit is true. The p-value is calculated as the proportion of the 160,000 null estimates at least as large in absolute value as the first estimate, obtained using the observed randomization and the FSF weights. For analysis 3, which employs IPWs, we pre-multiply the FSF and replicate weights by the IPW, and then repeat the same set of steps as described for analyses 1 and 2.

5 Appendix

The appendix is organized as follows:

- Section 5.1 describes the methods used for generating propensity scores that are used in the propensity-determined condition. These scores were generated regardless of a respondent's allocation to that condition.
- Section 5.2 describes the randomization process in greater detail.
- Section 5.3 describes the process for constructing the nonresponse adjustment factor (NAF) weights that we use for our exploratory analysis of the impact of T on effective sample size.

5.1 Propensity estimation procedure

Here, we outline the procedure we used to estimate the propensity scores. These were generated in winter of 2020 prior to the AHS 2021 fielding.

We used the following general process to (1) train the model, (2) validate the model, and (3) select the best-performing model.

1. **Begin with the "long-form" AHS data where each unit is repeated across four waves:** we observe response outcomes for the 2015, 2017, and 2019 waves; we are trying to predict response outcomes for the 2021 wave:

id	wave	respond	contact attempts	acs % white	...
1	2015	1	2	40	
1	2017	0	15	42	
1	2019	1	1	45	
1	2021	?			
2	2015	1	1	10	
2	2017	1	1	11	
2	2019	1	2	10	
2	2021	?			
3	2015	0	10	80	
3	2017	1	5	80	
3	2019	0	3	81	
3	2021	?			
⋮					

2. **For features, pull out the 2015 and 2017 waves and aggregate values so that each unit has one row:** we auto-generated three aggregations of either numeric or dummified variables: min (for the 0, 1 dummies, whether ever 0), max (for the 0, 1 dummies whether ever 1), and mean (for the 0, 1 dummies, percent 1). Auto-removal of highly-correlated features using `Caret` often removed the max and min, so we retained the mean except for the explicit lagged nonresponse features:

Feature matrix:

id	wave	mean contact attempts	mean acs % white
1	2015/2017	8.5	41
2	2015/2017	1	10.5
3	2015/2017	7.5	80
⋮			

3. **Augment that 2015/2017 feature matrix with the unit's response status in the 2019 wave**

Feature matrix with label:

id	wave	mean contact attempts	mean acs % white	2019 response
1	2015/2017	8.5	41	1
2	2015/2017	1	10.5	1
3	2015/2017	7.5	80	0
⋮				

4. **With this combined feature/label matrix, split the data into an (1) 80% training set (with five folds then used for cross-validation to tune hyperparameters) and (2) 20% validation set:**

Feature matrix with label and train/test status:

id	wave	mean contact attempts	mean acs % white	2019 response	Status	Fold
1	2015/2017	8.5	41	1	Train	1
2	2015/2017	1	10.5	1	Test	NA
3	2015/2017	7.5	80	0	Train	4
⋮						

5. **After estimating/tuning the models in the training set, evaluate the accuracy in the 20% held-out set using the metrics we discuss in Section 5.1.6**

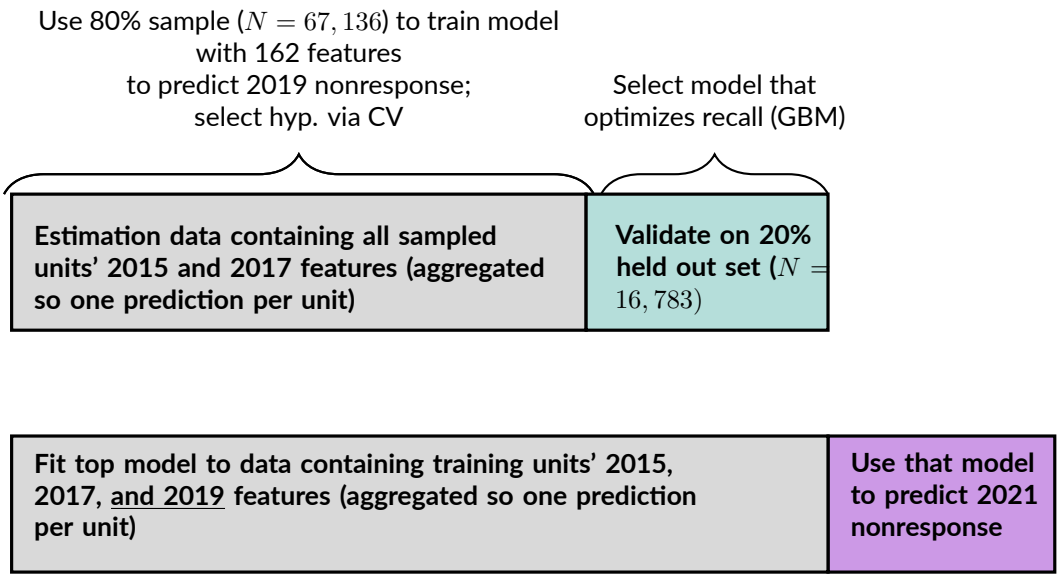
6. **Finally, for all units and (1) using the best performing model, and (2) an updated feature matrix to which the 2019 values are added/aggregated, predict nonresponse in 2021 to generate the $\hat{\eta}_i$ used in the field experiment:**

Feature matrix now including 2015, 2017, *and* 2019 and predicting 2021 nonresponse :

id	wave	mean contact attempts	mean acs % white	Predicted 2021 nonresponse
1	2015/2017/2019	6	42.3	0.22
2	2015/2017/2019	1.3	10.3	0.43
3	2015/2017/2019	6	80.3	0.87
⋮				

Figure A1 summarizes this process visually.

Figure A1: Process for prediction and validation



Here, we provide additional details on each step.

5.1.1 Label definition

In turn, there are a variety of outcomes we could predict corresponding to different types of nonresponse.

Noninterviews are split into three types:

1. **Type A noninterviews (focus of our prediction):** these occur when a regular occupied interview or usual residence elsewhere interview fails, usually because the respondent refuses, is temporarily absent, cannot be located, or presents other obstacles (such as language barriers the field staff are unable to overcome).
2. **Type B and Type C noninterviews:** each of these pertain to failures to interview someone about a vacant unit. If units are ineligible for a vacant interview during the attempt, but may be eligible later, they are classified as Type B noninterviews—for example, sites that are under or awaiting construction, are unoccupied and reserved for mobile homes, or are occupied in some prohibited manner. Type C noninterviews are ineligible for a vacant interview and will remain so, for example, because they have been demolished or removed from the sample.

Because we were focused on nonresponse to target *person-directed incentives*, which do not correspond to vacant interviews, we define the primary label as follows:

- `Nonresponder`: unit is a type A noninterview for any reason (so not only includes refusal but also not at home, language issues, etc.)
- `Others`: unit is either a responder (the vast majority) or a vacant noninterview.

5.1.2 Prediction methods we compared

5.1.3 Flexible binary classifiers

We fit a series of binary classifiers shown in Figure A3. With the exception of the neural network, the classifiers are *tree-based classifiers*. At its core, a tree-based classifier is an algorithm that is looking

to find combinations of attributes within which there are *only* responders or *only* nonresponders. Starting with the simplest version—a decision tree—imagine we start with two features: the Census region in which a unit is located and the percentage of households with a high school education or less. The classifier might first find that areas where fewer than 10 percent of households have HS education or less have units that are more likely to respond, creating a split at that value. The "tree" has its first "branch," with one group of people at the end of the "fewer than 10 percent" fork and another group of people at the "greater than 10 percent" fork. Now suppose that, among the first group, one region had proportionally many more responders than the other, but among the second group, region does not seem to make a difference. In that case, there will be a second branch between high- and low-responding regions among those in areas where fewer than 10 percent of people have a HS diploma, but no such split among those who live in the areas with more than 10 percent of people with HS diplomas. The maximum depth parameter constrains the number of splits and branches our tree can have.

Chance variation can lead to very idiosyncratic trees—the classifier tends to "overfit" to the data, meaning that its particular set of branches and splits will not do a good job of sorting responders from nonresponders in other samples. Random forest models (`RF`) are a solution to this problem that generalize the idea of decision trees. The idea is to fit many hundreds of decision trees (a forest) using two sources of random variation. One is random samples of the data with replacement; another is random subsets of the features used for prediction—so, for instance, rather than including all ACS features in a particular tree, one tree might have percent renters and racial demographics; another percent owners and racial demographics.

Finally, we employ gradient-boosting models (`GBM`). This is an *ensemble classifier*—each takes a series of shallow decision trees ("weak learners"). Adaptive boosting starts with a weak learner and then improves predictions over iterations by successively upweighting observations that were poorly predicted in iteration $i-1$. Gradient boost operates similarly, though instead of *upweighting* poorly predicted observations, it uses residuals from the previous iteration in the new model.

Overall, these tree-based classifiers aim to improve prediction by splitting and combining predictors. They generate what are called *feature importances*—measures of whether a predictor improves prediction of nonresponse. Importantly, feature importance metrics are directionless: that is, they measures how high up in a tree or how frequently an attribute is chosen, for example, irrespective of the sign or size of the coefficient.

We tuned the hyperparameters for these estimates using within-training set 5-fold cross-validation based on the eventual accuracy metric we focus on: recall.

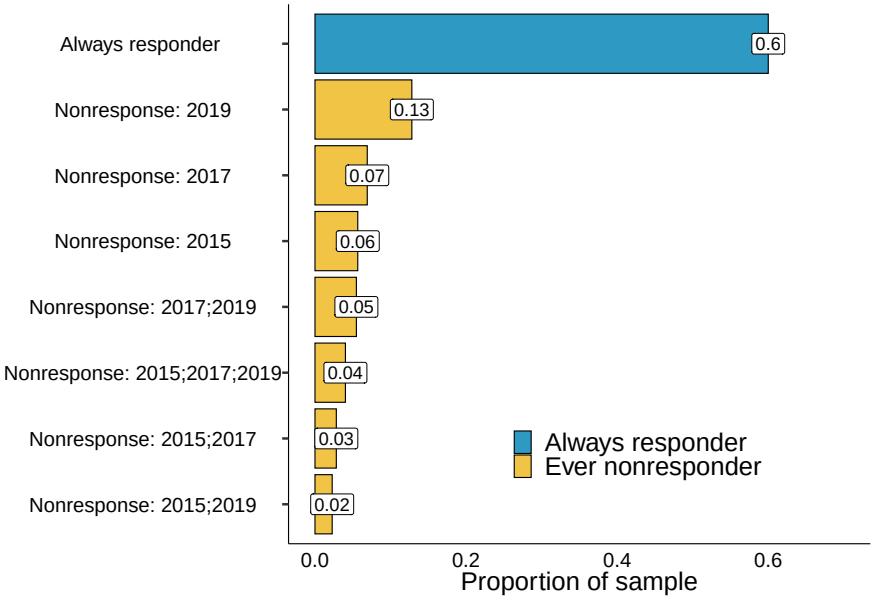
5.1.4 Baseline predictions

We then compared these flexible classifiers to two baseline methods:

1. **Empirically-informed guess:** with this method, we predict a unit is a nonresponder with probability equal to the empirically-observed proportion of nonresponders ($\sim$ 25% in the 2019 wave).

2. **Simple rule based on past nonresponse:** Figure A2 shows that past nonresponse behavior can be predictive of response behavior in a focal wave — for instance, 4% of units are never responders, and over 10% of units were nonresponders in two or more waves. In this baseline comparison, we use a simple rule where we predict a unit is a nonresponder if they were a nonresponder in the previous wave.

Figure A2: Response patterns across waves The figure focuses on Type A nonresponse/response that refers to behaviorally-driven nonresponse. The sample contains approximately 84,000 occupied units that were sampled for the panel starting in the 2015 AHS wave.



5.1.5 Predictors

We fit these models to two sets of features, imputing missingness to the modal value for categorical and mean for numeric:

1. `AHS-only` features from two sources:

   (a) **AHS sampling frame or master file variables**. We use binary indicators created from categorical levels of the variables that include the following:

      i. `DEGREE`: this is a measure of area-level temperature, and reflects places with hot temperatures, cold temperatures, and mild temperatures based on the number of heating/cooling days.

      ii. `HUDADMIN`: this is a categorical variable based on HUD administrative data for a type of HUD subsidy such as public housing or a voucher.

      iii. `METRO`: this is a categorical variable for the type of metropolitan area the unit is located in (e.g., metro versus micropolitan) based on OMB definitions for 2013 metro areas.

      iv. `UASIZE`: this is a categorical variable for different sizes of urban areas when applicable.

      v. `WPSUSTRAT`: this is a categorical variable for the primary sampling unit strata.

   (b) **Response and contact attempt variables from the previous waves**. We exploit the longitudinal nature of the data and use the unit's past response-related outcome to predict its status in a focal wave:

      i. total prior contact attempts (a numeric measure);

ii. the total number of interviews in the prior wave (capturing respondents who needed multiple interviews to complete participation);

iii. whether the unit was a nonresponder in the previous wave (binary).

2. `AHS + ACS` adds the following to the previous list:

   (a) **American Community Survey (ACS) 5-year estimates of characteristics of the unit's Census tract**. We list these variables in Appendix Table **??**. They were matched to waves as follows so that the predictor is measured temporally prior to the outcome: 2015 wave (ACS 5-year estimates 2009-2014); 2017 wave (ACS 5-year estimates 2011-2016); 2019 wave (ACS 5-year estimates 2013-2018). They reflect race/ethnicity, educational attainment, and different housing-related measures.

In the final model, we used the combined feature set.

After (1) dummifying all categorical variables, and (2) filtering out highly-correlated predictors in the estimation set, we ended up with 287 predictors. These predictors were:

1. **Aggregations across the 2015 and 2017 waves for the purpose of predicting 2019 response to select a best-performing model**
2. **Aggregations across the 2015, 2017, and 2019 waves for the purpose of predicting 2021 response, the propensities we use for our field experiment testing targeting**

5.1.6 Accuracy metrics

Finally, we evaluated the models in the held-out 20% data, with labels taken from the year 2019 (with features only corresponding to the 2015/2017 waves to avoid "leakage" of future knowledge into model estimation).

We examined three different outcomes of the predictions to calculate three separate evaluation metrics in the held-out or test fold. These are based on comparing a unit's actual nonresponse status to its predicted nonresponse status. Units can fall into four mutually exclusive categories, and the evaluation metrics are different summary measures of the categories across the entire held-out fold:

1. TP: a nonresponder is correctly predicted to be a nonresponder
2. FP: a responder is incorrectly predicted to be a nonresponder
3. FN: a nonresponder is incorrectly predicted to be a responder.[11]

From there, we constructed three composite measures as ratios of the total number of units falling into each category:

1. **Precision**: $\frac{\text{Total TP}}{\text{Total TP} + \text{Total FP}}$ Among predictions of nonresponders, what proportion are actually nonresponders;
2. **Recall**: $\frac{\text{Total TP}}{\text{Total TP} + \text{Total FN}}$ Among actual nonresponders, what proportion do we correctly predict to be nonresponders, as opposed to erroneously predicting that they are responders;

11. We do not need the fourth possible outcome of true negatives (correctly predicted responders), since $TN = 1 - TP - FN - FT$.

3. **F1 Score**: $2 * \frac{Precision * Recall}{Precision + Recall}$ Explained below.

If we have precision of 1, that means every time the model predicted a unit was a nonresponder, it actually was. For example, if there are 50 nonresponders and 50 responders, as long as the model predicts at least one nonresponder and no responders are falsely predicted to be nonresponders, it will have precision of 1. If instead, every time the model predicts a nonresponder that unit is actually a responder, its precision will be 0.

For recall, we have to look at the subset of *actual* nonresponders. If there are two nonresponders in a sample of 100 people, and the model predicts every single person in the sample is a nonresponder, then 100 percent of nonresponders are correctly predicted to be nonresponders and the recall will be 1. However, if the model does not predict any nonresponders to be nonresponders, its recall will be 0.

We used the F1 Score as a third summary metric, since it helps us balance between finding all nonresponders (high recall) while still ensuring that the model accurately separates out responders from nonresponders (precision). Note that one measure may be more useful over another in other applications. For an intervention targeting nonresponse bias, where there could be a higher cost to failing to predict nonresponse (false negatives) than to wrongly predicting nonresponse (false positives), we may prioritize models with high recall.

While what counts as a "good" F1 Score varies based on the context, generally, scores above 0.7 are considered evidence of a high-performing model. To gain more intuition, consider the simplified example in Table A1 of predictions for 20 units and where we use 0.75 as the cutoff for translating a continuous predicted probability of nonresponse (NR) to a binary label of NR or respond (R). Our precision is $\frac{3}{3+1} = 0.75$ since we have three true positives and one false positive. We could increase our precision through raising the threshold for what counts as a true predicted nonresponse to 0.8. However, doing so would hurt our recall which in the case of the example is $\frac{3}{3+3} = 0.5$ due to the presence of false negatives in the lower predicted probability range. The F1 Score is less interpretable than either of these since it combines the two, but in this case, it would be $2 * \frac{0.75 * 0.5}{0.75 + 0.5} = 0.6$, which is lower than what we observed in our real results. The example also shows that we can target our desired metric—for instance, capturing all nonresponders even if it leads to some false positives—by changing the threshold for translating a continuous value (e.g., $\hat{y} = 0.8$) into a binary prediction of nonresponse.

For the purposes of our field experiment, we selected the best model using `recall`. Our rationale is that, for the purpose of targeting incentives, we want to focus more on minimization of false negatives—respondents we fail to provide incentives to but who might be moved by that incentive to respond—than on wasted incentives on false positives—units that would have responded anyways.

Table A1: Illustration of the evaluation metrics: example predictions

ID	Pred. $\hat{y}$ continuous	Pred. $\hat{y}$ binary	True y	error_category
1537	0.99	NR	NR	True pos.
1177	0.93	NR	NR	True pos.
1879	0.84	NR	NR	True pos.
1005	0.78	NR	R	False pos.
1187	0.72	R	R	True neg.
1034	0.71	R	R	True neg.
1159	0.60	R	NR	False neg.
1181	0.52	R	NR	False neg.
1071	0.49	R	R	True neg.
1082	0.47	R	R	True neg.
1603	0.44	R	R	True neg.
1762	0.33	R	R	True neg.
1319	0.29	R	R	True neg.
1359	0.24	R	NR	False neg.
1238	0.21	R	R	True neg.
1490	0.17	R	R	True neg.
1465	0.17	R	R	True neg.
1338	0.11	R	R	True neg.
1766	0.07	R	R	True neg.
1807	0.04	R	R	True neg.

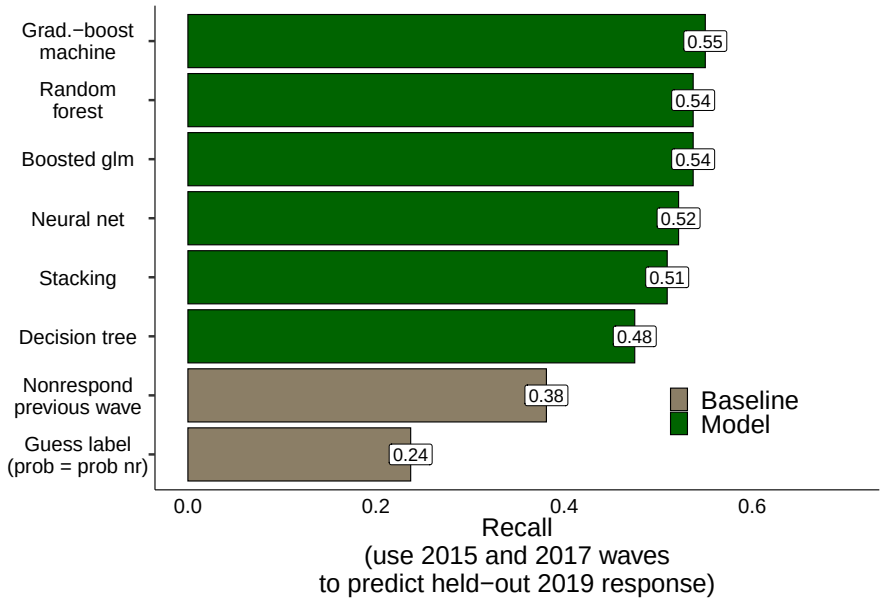
5.1.7 Results and model selection

Figure A3 shows the comparative accuracy of the two types of models: (1) flexible classifiers that predict behaviorally-driven nonresponse[12] and (2) baseline measures that correspond to the *status quo* methods survey planners might use.

The graph shows that large gains in prediction come from the move from (1) *no targeting* (or just guessing nonresponse status based on its empirical proportion) to (2) even simple, rule-based targeting of using the nonresponse status in the previous wave to assume persistence in that behavior. The least useful model (decision tree), which corresponds most closely to a rule-based approach but with model-selected splits on features rather than the researcher-selected feature of nonresponse status, still substantially outperforms that simple rule (a gain in recall of 10 percentage points, or a 26.3% improvement over the baseline rate). We then see a series of models with smaller variations in predictive accuracy. The best-performing model is the ensemble classifier of a gradient-boosting machine (`GBM`), but random forest also performs well. These two models provide a 17 percentage point improvement in recall over a simple rule of persistent behavior across waves, representing a 44.7% improvement over that rule-based baseline.

12. In the remainder, we use the terms nonresponse and behaviorally-driven nonresponse interchangeably to refer to the Type A nonresponse we discuss in Section 5.1.1.

Figure A3: Comparative accuracy in 20% held-out set between two ways to predict nonresponse and target incentives: (1) baseline targeting methods (random targeting; simple decision rule based on history of nonresponse), (2) risk-based targeting (either using a single model or ensemble method) The figure shows recall metrics in the 20% held-out set, with all features/predictors measured in 2015 and 2017 and the label corresponding to 2019. We focus on recall because the goal of risk-based targeting is to provide incentives to all potential respondents who may be swayed by those incentives, and we are more concerned with minimizing false negatives (finding all potential nonrespondents) than minimizing false positives (wasted incentives). This prioritization of recall over precision/other metrics may vary based on the size of incentive targeted.



Focusing on `GBM`, the best-performing model, useful is to examine how the metric of *recall* (1) breaks down into different categories of errors and (2) compares to the simple, rule-based prediction of previous nonresponse status. We see two observations:

1. **Why GBM performs better than that rule:** `GBM` is more accurately able to mitigate against false positives. Most notably, when we assume that a respondent's previous response status persists into the next wave, we have substantially higher rates of people who we predict as nonresponders but who actually respond (representing potentially-wasted incentives in the targeting framework). The more flexible classifier that weights not only nonresponse history but other attributes is better able to mitigate these false positives.
2. **How GBM could be improved:** conversely, `GBM`'s metric suffered from false negatives, or being overly optimistic in predicting who would respond. As Figure A5 shows, this likely also stems from the shifting base rates of nonresponse over time, with a relatively sharp increase in the 2019 wave (used to form the label) relative to the other waves.

Figure A4: Algorithm versus simple decision rule: types of errors Focusing on the best-performing algorithm, `GBM`, we compare the types of errors the algorithm makes to errors from the rule-based approach of previous nonresponse status.

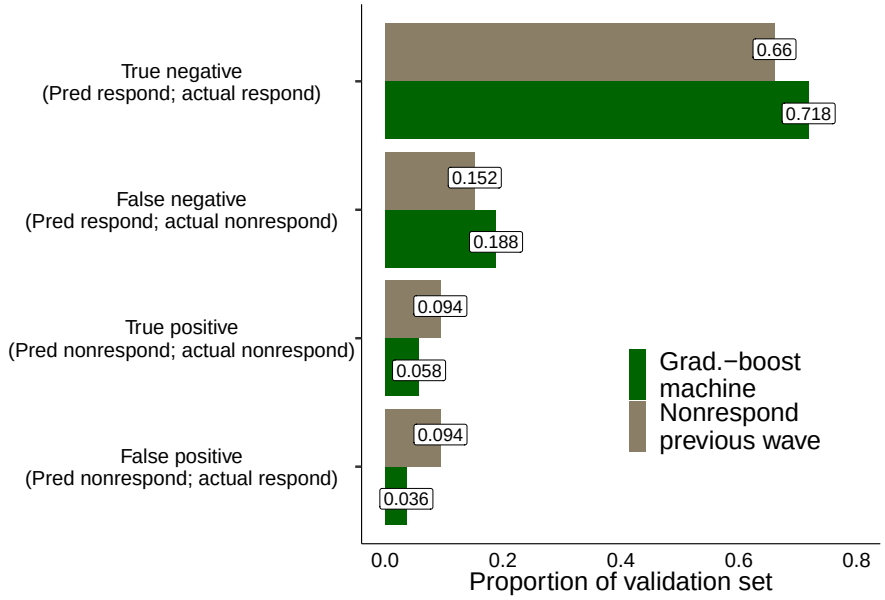
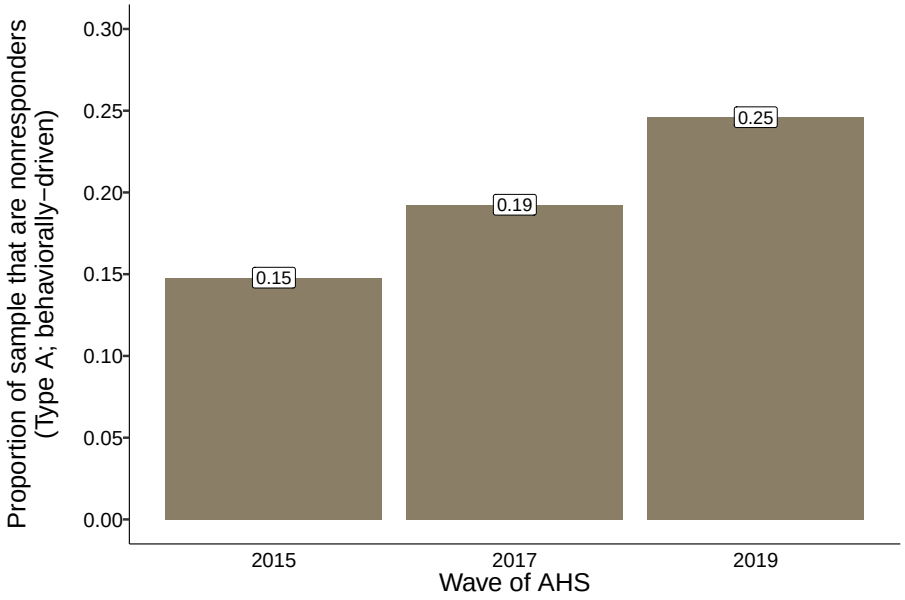


Figure A5: Possible source of false negatives in model— rising base rate of nonresponse The figure, again focusing on Type A, behaviorally-driven nonresponse, shows how the leap in base rates may drive false negatives, or model-predicted responders who go on to nonrespond.



Ultimately, given the superior performance of `GBM` relative to the other binary classifiers and to a simple, rule-based approach, we used `GBM` to generate the final predictions.

5.2 Randomization procedure

There are three variables that are randomly assigned: $T_i \in \{0, 1\}$ is an indicator for whether the unit receives the allocation they would have received under the Propensity-Determined (versus Propensity-Independent) method; $Z_i \in \{0, 1\}$ is an indicator for whether the individual is assigned to receive any incentive amount in the allocation used; $A_i \in \{0, 2, 5, 10\}$ is the dollar amount allocated to each potential respondent. The procedure for the random assignment works as follows, with $\hat{\eta}_i$ referring to the 2021 nonresponse propensities estimated in the previous section:

1. **Create** $Z_i^{T=1}$. Order each potential respondent from highest to lowest $\hat{\eta}_i$. Calculate $m \approx .3 \times n$, and assign the first $m - n$ individuals to $Z_i^{T=1} = 0$ and the last m to $Z_i^{T=1} = 1$. This provides the vector $Z^{T=1}$: the assignment that would have obtained, had each unit been assigned using Propensity-Determined Allocation.

2. **Create** $Z_i^{T=0}$. Define $f()$ as a function that randomly sorts a vector, and set $Z_i^{T=0} = f(Z_i^{T=1})$. This provides the vector $Z^{T=0}$: it is the assignment that would have obtained, had each unit been assigned to incentives using Propensity-Independent Allocation.

3. **Create** T_i. Sort individuals in order of their estimated propensity (randomly re-sorting within equal propensities) and form them into consecutive pairs. Within each pair, assign one individual to $T_i = 1$ and one to $T_i = 0$ with .5 probability. If there is an odd number of individuals, randomize the last unit using a coin flip.

4. **Create** Z_i. For all units for whom $T_i = 1$, set $Z_i = Z_i^{T=1}$, and for those for whom $T_i = 0$, set $Z_i = Z_i^{T=0}$.

5. **Create** A_i. Among units where $Z_i = 1$, randomly assign 50% to $A_i = 10$, 25% to $A_i = 5$, and 25% to $A_i = 2$. Assign the remaining sample for whom $Z_i = 0$ to $A_i = 0$.

5.3 Constructing weights to adjust for nonresponse

For the exploratory analysis of the impact on variance discussed in Section 4.9.1, we will construct weights that adjust for nonresponse separately for the treatment and control group. To do so, we will mimic part of the procedure AHS uses for its own weights construction, using the 2019 procedure (U.S. Census Bureau and Department of Housing and Urban Development 2020). We will:

1. Use the following variables to define discrete cells:
   (a) AHS administrative region (6 values)
   (b) Interview mode: in person; not in person[13]
   (c) Type of housing unit: house/apartment/flat; mobile home; other
   (d) 2013 metropolitan area at the county level: principal city; nonprincipal city; micropolitan area; non-CBSA area
   (e) Quartiles of census block group median income

Within the cells defined by the same variables, the noninterview adjustment factor (NAF) within each cell is defined as:

$$NAF = \frac{Interviews + noninterviews}{Interviews}$$

13. Question for Census: how is this defined for noninterviews?

Then, cells are collapsed if they either contain fewer than 25 units or have an NAF > 2.

References

Hansen, Ben B, and Jake Bowers. 2008. "Covariate balance in simple, stratified and clustered comparative studies." *Statistical Science*, 219–236.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. "Metalearners for estimating heterogeneous treatment effects using machine learning." *Proceedings of the national academy of sciences* 116 (10): 4156–4165.

Künzel, Sören R, Simon JS Walter, and Jasjeet S Sekhon. 2019. "Causaltoolbox—estimator stability for heterogeneous treatment effects." *Observational Studies* 5 (2): 105–117.

U.S. Census Bureau and Department of Housing and Urban Development. 2018. *2015 AHS Integrated National Sample: Sample Design, Weighting, and Error Estimation*. Technical report. https://www2.census.gov/programs-surveys/ahs/2015/.

———. 2020. *2019 AHS Integrated National Sample: Sample Design, Weighting, and Error Estimation*. Technical report. https://www2.census.gov/programs-surveys/ahs/2019/2019%20AHS%20National%20Sample%20Design,%20Weighting,%20and%20Error%20Estimation.pdf.

Wager, Stefan, and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.