Using Incentives to Reduce Nonresponse Bias in the American Housing Survey

U.S. Department of Housing and Urban Development | Office of Policy Development and Research

DISCLAIMER
The contents of this report are the views of the author and do not necessarily reflect the views or policies of the U.S. Department of Housing and Urban Development or the U.S. Government.

Using Incentives to Reduce Nonresponse Bias in the American Housing Survey

Prepared for

U.S. Department of Housing and Urban Development

Office of Policy Development and Research

Prepared by

Office of Evaluation Sciences

U.S. General Services Administration

September 2023

Acknowledgments

All acknowledgements are listed in alphabetical order. The authors of this report, Jasper Cooper, Michael DiDomenico, and Rebecca Johnson, would like to acknowledge several people for their valuable contributions. For contributions to the early ideas and design of the intervention and randomization: Melissa Cidade and Anne Herlache. For extensive feedback and support with data access: Shawn Bucholtz, George Carter, Tamara Cole, Sydney England, Emily Molfino, and Denise Pepe. For valuable feedback and support with approval: Kelly Bidwell, Amira Boland, Margo Schwab, Robert Sivinski, and Rita Young.

The U.S. Census Bureau reviewed this data product for unauthorized disclosure of confidential information and approved the disclosure avoidance practices applied to this release. Data Management System (DMS) number: P-1234567, Disclosure Review Board (DRB) approval number: CBDRB-FY22-349).

Table of Contents

List of Tables and Figures

Executive Summary

The American Housing Survey (AHS) is a biannual, longitudinal survey of housing units designed by the U.S. Department of Housing and Urban Development (HUD) and administered by the U.S. Census Bureau. As with many statistical surveys, the AHS has experienced declining response rates, raising concerns about nonresponse bias—divergence between estimates from the survey and the true value created by differences between respondents and nonrespondents. The present evaluation evaluates whether targeted cash incentives increase both response rates and decrease nonresponse bias.

In collaboration with HUD and the U.S. Census Bureau, the Office of Evaluation Sciences (OES) at the U.S. General Services Administration evaluated the effectiveness of targeting incentives during the 2021 wave of the AHS. The OES first estimated the risk of 2021 nonresponse for 86,000 potential respondents. Potential respondents with similar estimated risks of nonresponse were put into pairs and each member was assigned to one of two different groups: (1) a "targeted" group (n = 43,000) of potential respondents who received incentives if they were among the 30 percent with the highest estimated risk of nonresponse; and (2) a "nontargeted" group (n = 43,000) of potential respondents, each of whom had a 30-percent probability of receiving an incentive regardless of their estimated nonresponse risk. Potential respondents receiving an incentive via either method were randomly assigned a cash amount of $2, $5, or $10, delivered in a letter sent to all units in the survey.

Response rates were measured using data on response status and efforts to contact potential respondents from the 2021 fielding of the AHS. To estimate changes in nonresponse bias, OES estimated the joint statistical significance of differences between respondents in the targeted and nontargeted groups across 10 key attributes. The response rate among the group that received targeted incentives was an estimated 0.7 percentage points higher than the response rate of 67.2 percent in the group that received incentives completely at random. This result is statistically significant ($p = 0.018$).

Regarding nonresponse bias, a test of the joint statistical significance of differences across the 10 key attributes produced a p-value of 0.42. Any small observed differences could thus have been produced by random chance. Targeting did not measurably improve or worsen nonresponse bias. In additional analyses, no statistically significant evidence suggests that incentives increased response rates or decreased nonresponse bias overall; nor did evidence show that the marginal effectiveness of incentives at increasing the response rate diminishes with additional incentives.

Overall, this study provides limited support for the idea that targeted incentives can improve survey quality. Targeting incentives to those predicted to be most at risk of nonresponse was more effective at increasing response rates than simply allocating incentives at random. Holding the total incentive budget constant, simply targeting incentives to the potential respondents with the 30-percent highest predicted risk of nonresponse induced an additional 300 people to respond, compared with allocating incentives totally at random; however, these improvements in response rates are small in magnitude (with less than one percentage-point improvement) and are not accompanied by discernable reductions in nonresponse bias. It should be noted that the data collection overlapped with the onset of the COVID-19 pandemic and associated lockdowns, which may have muted the effectiveness of financial incentives.

Introduction

The American Housing Survey (AHS) is a biannual, longitudinal survey of housing units designed by the U.S. Department of Housing and Urban Development (HUD) and administered by the U.S. Census Bureau. The housing unit sample, drawn from the Census Master Address File, is designed to provide statistics representing both the entire country and its largest metropolitan areas. In 2015, and for the first time since 1985, a new sample was drawn, and new households were asked to participate in the survey.

As with many federal surveys, the AHS has experienced declining response rates, requiring increasing time and effort to reach the 80-percent response rate the Office of Management and Budget prefers. Response rates have declined, starting at approximately 85 percent in the 2015 wave, declining to 80.4 percent in the 2017 wave, and falling to 74.3 percent in the 2019 wave.

As response rates decline, data quality issues become increasingly important. This project experimentally evaluates the use of targeted monetary incentives to improve AHS data quality, learn which incentive allocation methods most effectively manage cost, and increase data quality, focusing on survey nonresponse bias.

Nonresponse bias is the divergence between a population quantity of interest and its sample estimate that is created by systematic differences between those who do and do not respond to a survey. For example, if nonresponders are more likely to live in inadequate housing than responders, an uncorrected sample estimate of housing inadequacy would bias lower than the true proportion of U.S. adults living in inadequate housing.

It is theoretically possible to adjust survey estimates to account for differential nonresponse so that those sample estimates converge to their corresponding population quantities, removing bias. The AHS calculates a noninterview adjustment factor that reweights for nonresponse within cells defined by the unit's metropolitan area, type of housing unit, block group median income, and area-level rural or urban status. Such adjustments, along with raking, may reduce nonresponse bias.[1] However, nothing guarantees that the model used for bias adjustment contains all the information it needs. Moreover, the weights used in such bias adjustment methods typically increase variance in estimates.

Consider the hypothetical population and sample in Table 1. The population is divided evenly into four strata: rural versus nonrural addresses and southern versus Midwestern addresses. A random sample of 100 respondents, 25 in each stratum, estimates the proportion of residents living in inadequate housing. The housing inadequacy rate is highest in the northern, nonrural stratum, which also has the lowest response rate, 10 / 25 = 40 percent versus 25 / 25 = 100 percent everywhere else. The true rate of housing inadequacy is (3 x 2500 x 8 percent + 2500 x 20 percent) / 10000 = 11.0 percent. If the uncorrected sample average is calculated, the answer is still biased downward: (3 x 25 x 8 percent + 10 x 20 percent) / 85 = 9.4 percent. To adjust for

[1] The AHS raking procedure, as implemented in the 2019 wave, is described in section 3.4 (U.S. Census Bureau and HUD, 2020). This process broadly involves using "control totals"—or known estimates of housing and population totals from other sources—to adjust the weights on AHS respondents, moving the AHS sample estimate of the housing or population characteristic closer to the control or independent estimate. The AHS defines a priority order for adjustment, because moving sample estimates closer to control or independent estimates on one attribute (for example, the number of vacant housing units in a state) can cause sample estimates to move *further* away from population estimates for other attributes (for example, the number of people aged 65 or older in a state).

nonresponse, each respondent in the sample can be reweighted by the number of people in the population whom they represent (underlined): (3 x 25 x 8 percent x 100 + 10 x 20 percent x 250) / (3 x 25 x 100 + 10 x 250) = 11.0 percent.

Table 1. A Hypothetical Example of Nonresponse Bias Adjustment

Region	Rural	% Housing Inadequate	Population (N)	Attempted Interviews	Successful Interviews	Population Represented by Respondents
south	Yes	8	2500	25	25	2500/25 = 100
south	No	8	2500	25	25	2500/25 = 100
north	Yes	8	2500	25	25	2500/25 = 100
north	No	20	2500	25	10	2500/10 = 250

Removing nonresponse bias comes at a cost in terms of variance, however. The uncorrected biased sample mean standard error is 3.2 percentage points, whereas the mean standard error reweighted to remove nonresponse bias is 3.4 percentage points.[2] The standard error had to increase by 7 percent to decrease nonresponse bias by 14 percent. Generally, correcting for nonresponse bias on the backend presents a bias-variance tradeoff, because the respondents in strata with low-response rates need to "represent" more nonrespondents than the respondents in strata with higher response rates. Alternatively, improving data quality on the front end by increasing response rates among those who would otherwise contribute to nonresponse bias can reduce both bias and variance in the resulting estimates. This study examines the potential for doing so using targeted monetary incentives.

Appendix A shows the results from an analysis of the characteristics of 2015 respondents who did not respond to the subsequent 2017 AHS wave. Systematic patterns emerge regarding who drops out from the panel. For example, units with younger householders, as reported in the 2015 AHS, were more likely to be nonresponders in the 2017 AHS than units with older householders, as reported in the 2015 AHS. Although these findings do not provide conclusive evidence that declining AHS response rates produced nonresponse bias, they do indicate attrition from the panel is correlated with key outcomes of interest, signaling the potential for data quality improvements in the AHS.[3]

The present project aims to determine whether and how providing cash incentives before Census Bureau contact reduces nonresponse bias in (adjusted and unadjusted) sample estimates by increasing the response rate among those deemed to be at the highest risk of nonresponse. Although providing large incentives to all housing units in the sample could conceivably increase both the response rate and data quality, the goal of the project is not to test the effectiveness of blanket incentives. Rather, the study is designed to generate evidence about the effectiveness of targeting incentives to different types of units, aiming to efficiently use incentives to convert the subset of important cases that would not participate in the survey otherwise. The goal is to move away from a uniform allocation of incentives, which is inefficient

[2] Standard errors of the weighted and unweighted means calculated by regressing a binary indicator for housing inadequacy among 85 hypothetical respondents on an intercept, with and without the weights in Table 1.
[3] The Nonresponse Bias Memo in appendix A is an analysis the Office of Evaluation Sciences completed and does not represent the official views of HUD nor the Census Bureau.

in providing incentives, both to cases unlikely to be affected by incentives and cases unlikely to introduce bias.

Overview of Experimental Design

The intervention sends different levels of cash incentives to potential respondents sampled as part of the 2021 American Housing Survey (AHS) Integrated National Sample. One-half of the housing units in the sample (the nontargeted group) were randomly selected to receive cash incentives. The other one-half were selected randomly to receive incentives based on the estimated likelihood of their nonresponse to the survey (the targeted group). The incentive level ($2, $5, or $10) was chosen completely randomly, contingent on being assigned to receive an incentive. The unselected housing units did not receive an incentive (no incentive group).

The cash was delivered inside an envelope containing a letter reminding the potential respondent about the survey (see appendix D). This letter was sent to all potential respondents, regardless of selection to receive incentives, albeit with slight wording changes mentioning the incentive in the letter sent to those potential respondents selected to receive incentives. The amount of cash in the envelope varied from $0 for unselected housing units to $2, $5, or $10 for those units selected to receive incentives.[4] Table 2 shows the distribution of different incentive amounts for the group with randomly allocated incentives and the group with targeted incentives. The sample sizes for the different incentive amounts were chosen to meet varying priorities. For budgetary reasons, no more than 30 percent of the sample received an incentive, and no incentive was greater than $10. The higher incentive amount of $10 was estimated to be the most likely effective, but researchers wanted to understand diminishing marginal returns. Therefore, to maximize the chances of finding any effect while enabling marginal return estimation, one-half of the 30 percent receiving incentives were allocated to the $10 group and one-fourth to the lower denominations of $2 and $5, respectively.

Table 2. Sample Size

Random Assignment of T (Targeting Method)							
Random Incentives (50%)				Targeted Incentives (50%)			
N = 43,000 (50%)				N = 43,000 (50%)			
Random Assignment of A (Dollar Incentive Received)							
$0	$2	$5	$10	$0	$2	$5	$10
30,000	3,200	3,200	6,500	30,000	3,200	3,200	6,500
70%	7.50%	7.50%	15%	70%	7.50%	7.50%	15%

Note: The counts are rounded according to U.S. Census Bureau entity count rounding rules.

Even if incentives increase the response rate, it does not necessarily imply a reduction in nonresponse bias (or increased data quality).[5] The effective and efficient use of incentives would target incentives only to those who convert to respondents (not those respondents who would respond, even in the absence of incentive or nonrespondents who would not respond, even with an incentive) and in the least amount able to convert them to a respondent, which varies depending on their propensity to respond.

[4] The bill dominations were one 2-dollar bill for the $2 treatment, one 5-dollar bill for the $5 treatment, and one 10-dollar bill for the $10 treatment.

[5] For a more comprehensive explanation of the motivation behind the study design and a review of relevant literature, refer to the Design Document in appendix B.

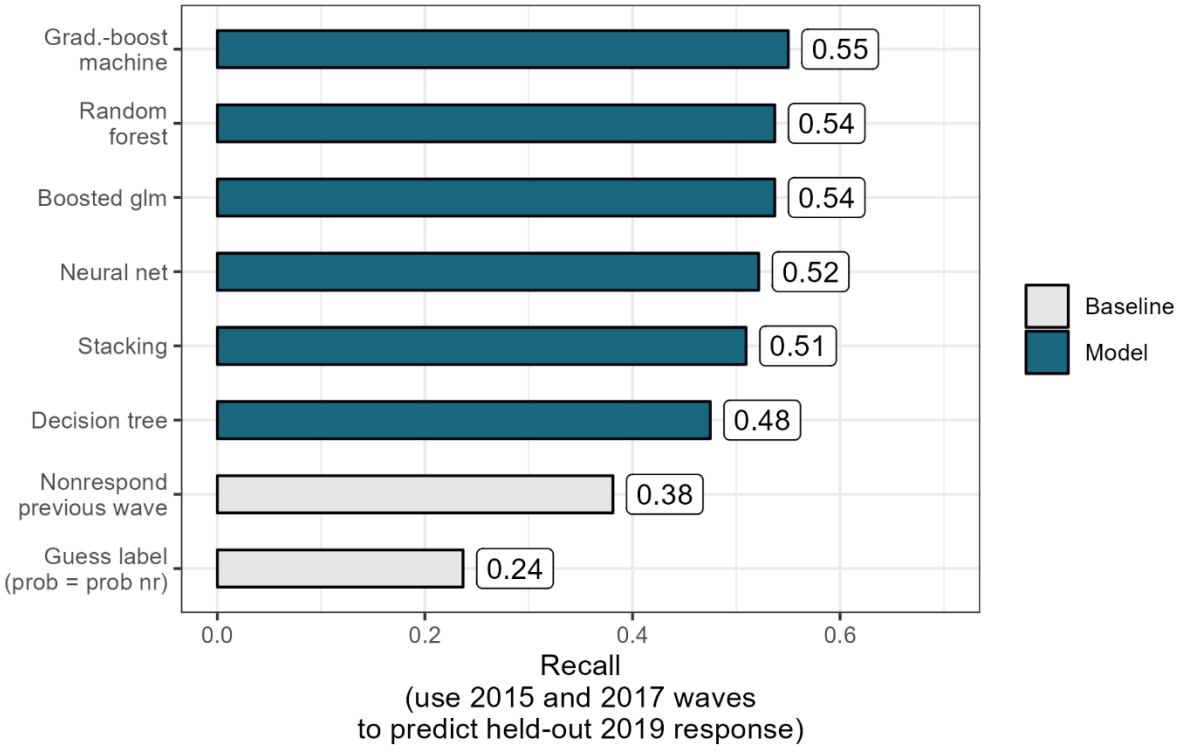
A key innovation of this study is the randomization of the *method* through which incentives were allocated. The project Design Document (appendix B) includes past literature reviews (section 2) and present study advancement discussions (section 2.4). This study estimates the causal effect of the targeting method on the sample composition and response rate by randomly dividing the sample into groups that receive targeted incentives and incentives without any targeting at all (completely random selection).

The targeting method was implemented using tools from supervised machine learning to predict a unit's nonresponse risk in 2021, given what is known about that unit from the sampling frame attributes and past waves. Appendix C provides a high-level description of nonresponse risk estimation.

The OES first developed a set of shared variables among the 2015, 2017, and 2019 AHS waves and merged those three waves. Eighty percent of the sample, selected at random, was used to train a series of flexible binary classifiers to predict 2019 nonresponse. The prediction algorithms included gradient-boosted machines, random forest, boosted generalized linear model, neural nets, regression tree, and a stacked ensemble. The model performance was tested using the remaining 20 percent of the data to prevent overfitting to the 2019 wave and to meet the goal of using the model to predict nonresponse in 2021. Specifically, machine learning models were compared with two benchmarks. The first performance benchmark randomly predicted 25 percent of the 2019 sample to be nonresponders (based on the nonresponse rate of 25 percent). The second benchmark simply labeled any 2015 or 2017 nonresponder as a predicted 2019 nonresponder. The performance of the machine learning models was compared with these benchmarks using recall as the performance metric. Recall measures the proportion of actual 2019 nonresponders correctly predicted to be nonresponders. Recall was chosen as the performance metric, because the goal of the incentive targeting is to provide incentives to all potential respondents whom such incentives may sway. This result requires a prediction algorithm well-suited to finding all potential nonrespondents.

Based on the results in Figure 1, the gradient-boosted machine method was selected, because it correctly predicted twice as many nonresponders than a random guess and (55 - 38) / 38 = 44 percent better than if one were to infer future response status using past response behavior. This model, fitted to the 2015–19 data, generated response predictions among the 2021 sample frame. In the one-half of the 2021 AHS sample randomly assigned to receive targeting, 30 percent of respondents estimated to be at the highest risk of nonresponse received incentives.

Figure 1. Actual Proportion of 2019 Nonresponders Predicted to Be Nonresponders Using Models Trained on Data From the 2015 and 2017 American Housing Surveys



The causal effect of using a targeted incentives strategy compared with a strategy in which incentives are allocated randomly without targeting can be estimated, because the very *method* used to allocate incentives is randomized (targeted versus random). This design is intended to generate insights for optimal incentive use.

This document refers to the variable that assigns respondents to either of the two incentive allocation methods as T, which stands for "targeting." When $T = 1$, the potential respondent receives the incentive allocation that they would receive if the targeted allocation strategy were used for the whole sample, and when $T = 0$, the potential respondent receives the incentive allocation that they would receive if incentives were assigned completely at random to the whole sample.

Potential respondents are randomly assigned to $2, $5, or $10, conditional to being allocated any incentive. This variable is denoted using A, for amount. The Analysis Plan (appendix C) describes the precise randomization procedure and resulting sample sizes.

Hypotheses

The research design is tailored to address a family of questions on how different kinds of incentive schemes affect data quality and the effort needed to collect the data. The following outlines the general sets of hypotheses and discusses the estimands for each hypothesis and estimation strategy in greater detail.

Primary Hypothesis: Impact on the Response Rate and Nonresponse Bias

Nonresponse bias in the American Housing Survey (AHS) is defined as a nonzero expected difference between unadjusted AHS sample estimates and their corresponding population

statistics (see the table 1 discussion, for example). Although nonresponse bias and the response rate are likely related, it is conceivable that incentives could change nonresponse bias independently of the response rate and vice versa. For example, if those induced to respond to the survey by incentives closely resemble the group that would have responded regardless of incentives, then the sample *composition*—the overall attributes of responders—would not change even though the response rate increased. Furthermore, if incentives do not work as intended and induce equal proportions of potential respondents to respond and not respond, the response rate may not change, even though the sample composition does. This second scenario also raises an important point for measuring nonresponse bias reduction: simply changing the sample composition is insufficient to infer a reduction in nonresponse bias. To cause a *reduction* in nonresponse bias, the change in the sample composition that incentives generate must bring the unadjusted sample estimates closer to their population targets. Knowing if changes brought about by incentives *improve* nonresponse bias in the sample composition requires comparing those changes in the sample composition with an unbiased benchmark. For this project, the ideal incentive intervention would increase the response rate and decrease the distance between unadjusted sample estimates and their corresponding quantities in the population. Relying on reweighting and increasing the amount of data available reduces sample estimates variance and potentially reduces bias not accounted for in adjustments in estimates.

The following describes a two-step procedure for detecting changes in nonresponse bias. First, the detection of *any* change in the sample composition is attempted. Second, and only if some change in the sample composition is detected, that change is compared with an unbiased benchmark. Given the potential independence of changes in the response rate from changes in nonresponse bias, two separate hypotheses are made:

1. Allocating the entire incentive budget to housing units deemed a higher nonresponse risk (targeted) **reduces nonresponse bias** compared with a purely random allocation of incentives (nontargeted).
2. Allocating the entire incentive budget to housing units deemed a higher nonresponse risk (targeted) **increases the response rate** compared with a purely random allocation of incentives (nontargeted).

Secondary Hypothesis: Impact on Required Effort

Incentives may decrease the overall effort required to collect the AHS data by inducing potential respondents to respond earlier. This study investigates the following hypotheses:

1. Providing any incentives (targeted and nontargeted) **reduces the number of contact attempts** Census enumerators undertake and **increases the response rate.**
2. Allocating the entire incentive budget to housing units deemed at higher risk of nonresponse (targeted) **reduces the number of contact attempts** Census enumerators undertake.

Tertiary Hypothesis: Diminishing Marginal Returns

The differing incentive amounts allocated allow for investigating the presence of "diminishing marginal returns," that is, the idea that an inflection point exists beyond which each additional dollar of incentive provided generates marginally less increase in the response rate. The hypothesis is formulated because the linear change in the response probability brought about by each additional dollar of incentives provided is lower at greater incentive amounts.

Statistical Models and Hypothesis Tests

Several variables are randomized in this study. Understanding the distributions of these variables is crucial to obtaining unbiased estimates of the treatment effects and uncertainty around these estimates.

Treatment Conditions and Probability Weights

Three variables are randomly assigned. Variable $T_i \in \{0,1\}$ indicates whether the unit receives the allocation it would have received under the targeted (versus randomly allocated) method, $Z_i \in \{0,1\}$ indicates whether the individual is assigned to receive any incentive amount in the allocation used, and $A_i \in \{0,2,5,10\}$ indicates the dollar amount allocated to each potential respondent. Respondents allocated $0 do not receive an incentive.

To randomly assign T, units are first formed into pairs with the most similar predicted nonresponse risk. Within each pair, one unit is randomly assigned to T = 1 and the other to T = 0. In this way, the random division of the sample into targeted and nontargeted groups is balanced with the estimated risk of nonresponse. These pairs constitute "pair blocks" and each has an individual identifier (ID).

The assignment procedure generates a correlation between the predicted probability of nonresponse and the probability of receiving an incentive, because incentives provided to the potential respondents are estimated to be in the top 30 percent highest risk of nonresponse wherever T is set at random to be 1. This correlation could cause bias if not accounted for, because it overrepresents certain covariate profiles and types of potential respondents in the group assigned to receive incentives. To correct this issue in any analyses assessing relationships between incentive receipt and outcomes, potential respondents who receive incentives are reweighted by the inverse of the probability that they are sent incentives and those who *do not* receive incentives by the inverse of the probability that they are not sent incentives.

It is possible to calculate both probabilities analytically using the probability that a given respondent is assigned to receive incentives. Because T is independent for any given individual, the probability of receiving incentives is given by—

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

The 30 percent of units with the highest estimated risk of nonresponse (allocated an incentive under targeting) evaluates to $0.5 \times 1 + 0.5 \times 0.3 = 0.65$. The 70 percent of units with the lowest estimated risk of nonresponse (not allocated an incentive under targeting) evaluates to $0.5 \times 0 + 0.5 \times 0.3 = 0.15$. Thus, four possible values exist for a treatment assignment probability $\pi_{i,z}^{Z}$ (where z indicates a treatment status for respondent i).

1. For j, the highest 30-percent risk of nonresponse individuals—
   a. Probability of receiving an incentive: $\pi_{j,1}^{Z}$ = 0.65.
   b. Probability of not receiving an incentive: $\pi_{j,0}^{Z}$ = 1 - 0.65 = 0.35.
2. For k, the lowest 70-percent risk of nonresponse individuals—
   a. Probability of receiving an incentive: $\pi_{k,1}^{Z}$ = 0.15.
   b. Probability of not receiving an incentive: $\pi_{k,0}^{Z}$ = 1 - 0.15 = 0.85.

Thus, even with deterministic incentive allocation when T = 1, observing every unit in every incentive condition is possible, albeit with differing probabilities. To obtain an unbiased estimator of the average treatment effect of receiving incentives, potential respondents overrepresented in incentive or no-incentive groups are downweighted, and those who are underrepresented are upweighted using $\frac{1}{\pi_{i,z}^{Z}}$. This adjustment is referred to as the inverse propensity weight (IPW) throughout.

Variance Estimation and Sample Weights

Given that this study involves a randomized experiment within a random survey sample, two potential sources of variation in estimates exist to account for when characterizing the statistical uncertainty around them: the random assignment of the experimental variables T, Z, and A and the random sampling of units from the sample framed into the sample. The approach taken for variance estimation depends on the underlying estimated effect—the average treatment effect of T, Z, or A in the specific *sample* obtained in practice, often called the sample average treatment effect (SATE); or the average treatment effect of those variables in the *population that the sample represents*, often called the population average treatment effect (PATE). To estimate the SATE, the random assignment of T, Z, and A must account for the resulting uncertainty. The data do not need reweighting to make the point estimates representative of the population. However, to estimate the PATE, the data must be reweighted to account for the survey sampling design (using the base weight that is not corrected for nonresponse bias) and employ variance estimators that account for variation resulting from both the random assignment and the random sampling procedures.

Our main analyses focus on whether incentives had the intended effect among *the units in the sample obtained in practice*. In other words, the interest is in the SATE. As such, the base sample weights are not employed in point estimates, and the variance estimation accounts only for the random assignment mechanism. Specifically, a p-value for the sharp null hypothesis of no treatment effect for any unit in the sample is obtained by conducting randomization inference (RI). This procedure involved randomly simulating T, Z, and A 5,000 times and reestimating the estimates to obtain their sampling distribution when the sharp null hypothesis is true. The p-value corresponds to the proportion of estimates obtained through this simulation procedure that is at least as large in absolute value as the observed estimate, referred to as the "RI p-value" throughout.

Estimates of the PATE serve as a robustness check for any statistically significant estimates of the SATE. To estimate the PATE, point estimates are reweighted by the product of the IPWs (when applicable) and the base sample weights and estimated RI p-values, including variance from Census-produced replicate weights corresponding to the sample base weight. For each of the RI simulations of T, Z, and A, 160 estimates are computed, each corresponding specifically to a different replicate weight. This robustness check is run on only statistically significant results, because the p-value is extremely unlikely to be smaller for the PATE than for the SATE. In general, including sampling variation from replicate weights into the null distribution generated through RI increases the sampling distribution's spread, increasing the probability mass in the distribution tails and, thereby, the p-value.

Overview of Statistical Analyses

This section presents five analyses. Three of these analyses are preregistered in the analysis plan included in appendix C, and the two analyses marked with asterisks (*) are not preregistered. The nonpreregistered analyses were conducted to add additional context to the differing incentive amount effects. The analyses are grouped based on the two treatments:

- Main randomized treatment of interest—targeting incentives versus random allocation, indicated using the variable T, or "Targeting." Analyses 1 through 3 focus on this treatment.
- Secondary randomized treatment of interest—receiving $0, $2, $5, or $10 of incentives, indicated using the variable A, or "Amount." Analyses 4 through 6 focus on this treatment.

Analysis 1: Impact of Targeting Incentives on Nonresponse Bias

This analysis focuses on key attributes of housing units, households, and areas the 2021 AHS measured. This list, developed based on the nonresponse bias analysis and conversations with the Census Bureau, includes the following variables from the AHS Internal User File (IUF) or sampling frame:[6]

1. Own house (no; yes, with mortgage or loan; yes, with no mortgage or loan).
2. Average household size.
3. White alone (householder).
4. Age of householder.
5. Presence of rodents.
6. Presence of mold in any room.
7. *Sampling frame:* Census Division.
8. *Sampling frame:* HUD-assisted unit (as of 2013).[7]
9. *Sampling frame:* 2013 metropolitan area (county-level; principal city, nonprincipal city, micropolitan area, noncore-based statistical area).
10. *Sampling frame:* Type of housing unit (house or apartment, mobile home, other).

Conducting individual tests for each outcome poses a multiple comparisons problem. Therefore, the analysis includes an omnibus test of the null hypothesis that the (conditional) differences in means are zero across all outcomes, including between the sample with targeted incentives and the sample with randomly assigned incentives. Specifically, an F-test compares a model in which the allocation method indicator, T, is regressed on the pair block ID variable with one in which T is regressed on the pair block ID variable and the list of outcomes, both described previously.

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 implies that at least one of the outcomes is imbalanced with T. Thus, it is possible to run one test to determine whether the first moments of the distributions of any outcomes differ between the two allocation methods. If the

[6] These variables derive from three sources: (1) variables from when the 2015 AHS estimates deviated significantly from 2010 Decennial Census estimates (figure 1 in nonresponse bias summary memo); (2) variables that are predictive of panel attrition using a large penalty term from a model (figure 15 in the nonresponse bias summary memo); and (3) sampling frame variables used in the nonresponse adjustment process.
[7] U.S. Department of Commerce and HUD (2015: 3).

null hypothesis is rejected—if at least one of the outcomes varies between the group randomized to targeted incentives and the group randomized to random incentives—should this change in the sample composition be interpreted as a reduction or an increase in nonresponse bias?

Answering this question depends on the differences in sample composition test findings.

- If no statistically significant differences exist in sample composition between the two groups, there is no evidence that the treatment decreases or increases nonresponse bias. If the targeting produces no statistically significant differences in sample composition between the two groups, it is not possible to find evidence of a reduction or increase in nonresponse bias, because the distance of each group's sample quantities from population-level targets is statistically indistinguishable.
- If there *are* statistically significant differences in sample composition between the two groups, nonresponse bias is investigated further using a benchmark data source. If the targeting produced statistically significant differences, the preanalysis plan (Appendix C) preregistered using a benchmark data source to graphically investigate which group's values were closer to benchmark population quantities. This result could explain whether the targeting decreased bias (values closer to those quantities) or increased bias.

The preanalysis plan in Appendix C contains additional details on the estimation procedures.

Analysis 2: Impact of Targeting Incentives on Response Rate and Enumerator Effort

This analysis focuses on whether targeting improves two outcomes.

1. **Response Rate.** This outcome is 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. **Enumerator (Field Representative) Effort.** This outcome measures the number of attempts survey enumerators (field representatives) made in an effort to complete a survey with a potential respondent. It is a continuous variable based on the Contact History Instrument, or CHI, data and aggregating contact attempts across all modes.[8]

The steps to estimate are—

- Regress each outcome on the pair block ID variable and T.
- Like the first analysis, IPWs are not used, because T is independent of units' covariates and potential outcomes.
- **For inference,** RI uses $m = 5{,}000$ replicates and a two-tailed p-value.
- In cases where a result is statistically significant, replicate weights, which account for the variance from the experimental design and the variance from the sampling procedure for including housing units in the AHS sample, are used as a robustness check.[9]

[8] Another measure of effort includes in-person contact attempts. The focus here is on all modes, because it reflects phone-based efforts, as well.

[9] Our analysis plan specified using the replicate weights for inference as a robustness check on both response rate and contact attempts. However, challenges with computational running time required limiting the robustness analysis to only significant results.

Analysis 3: Impact of Incentives on Nonresponse Bias

This analysis is essentially the same as Analysis 1, with two main changes. F-tests compare a restricted and a full model as previously done but rely on an indicator for receiving any incentive, Z, as the outcome in place of the indicator for targeting, T. IPWs reweight both regressions as described previously.

Analysis 4: Impact of Incentives on Response Rate and Enumerator Effort

To contextualize the results, particularly the analysis of marginal returns, understanding the overall effect of receiving incentives on the response rate and enumerator effort is useful. Two additional tests are described here, neither of which were preregistered.

1. The average effect of receiving $2, $5, or $10 incentives was estimated by comparing the response rates and contact attempts of those who received $0 with those who received $2, $5, or $10, respectively.
2. The average per-dollar linear change in response probabilities or contact attempts was estimated from $0 all the way to $10 incentives.

The estimation strategy in all such analyses features a linear regression among all potential respondents in the experimental sample, weighted by the IPWs. Each model is estimated on the two previously described outcomes—a binary indicator for response or an integer count of contact attempts. The ID variable for pair blocks is included in the regressions throughout. Each test specifies the incentive amount variable differently. The first test splits the incentive allocation variable, A, into three binary indicators, receiving $2, $5, and $10. Each coefficient on the incentive amounts—denoted as A2, A5, and A10—indicates the effect of receiving $2 versus $0, $5 versus $0, and $10 versus $0. The second test specifies A as a continuous variable, running from $0 to $10. The coefficient on this variable, thus, captures the average linear change in response rates and contact attempts for each additional incentive dollar, restricting the marginal effect to be the constant across this range. The next analysis relaxes this restriction.

Analysis 5: Diminishing Marginal Returns to Incentives

This analysis tests for the presence of diminishing marginal returns in the relationship among dollars of incentives provided, probability of response, and number of contact attempts.

Potential respondents are assigned with unequal probabilities accounted for through the IPWs to four different incentive amounts: A = 0, A = 2, A = 5, and A = 10. Diminishing marginal returns require first estimating the marginal returns, defined as the estimated linear response rate change and contact attempts for each incentive dollar spent.

The regressions in Analysis 4, the response rate and contact attempt outcomes on binary indicators for A = 2, A = 5, and A = 10 (and block indicators), estimate these marginal returns. A2, A5, and A10 denote the coefficients from this regression, which measure the estimated differences among receiving $0 and $2, $0 and $5, and $0 and $10. These coefficients estimate three marginal changes.

1. “A = 0 to A = 2”: the average, linear, per-dollar change in the outcome between $0 and $2, estimated using A2 / 2.
2. “A = 2 to A = 5”: the average, linear, per-dollar change in the outcome between $2 and $5, estimated using (A5 - A2) / 3.

3. "A = 5 to A = 10": the average, linear, per-dollar change in the outcome between $5 and $10, estimated using (A10 - A5) / 5.

Diminishing marginal returns are calculated by taking the differences in these marginal changes. Specifically, diminishing marginal returns are evaluated in two ways, as preregistered:

1. Subtract "A = 0 to A = 2" from "A = 2 to A = 5" to estimate the difference in marginal returns between $2 and $5 versus $0 and $2.
2. Subtract "A = 2 to A = 5" from "A = 10 to A = 5" to estimate the difference in marginal returns between $5 and $10 versus $2 and $5.

The first expression, for example, estimates the difference that an additional incentive dollar makes when moving from $2 to $5 versus when moving from $0 to $2. As a hypothetical example of diminishing returns, a value of -0.005 for the response rate outcome indicates that each additional incentive dollar spent when moving from $2 to $5, on average, produces ½ a percentage point less of an increase in the response rate than the average additional dollar spent when moving from $0 to $2. Diminishing marginal returns for the contact attempts variable are signed in the opposite direction. For example, a positive value of 0.005 from the first expression indicates that each additional incentive dollar spent when moving from $2 to $5 is 0.005 contact attempts *less* effective at reducing enumerator effort than the average additional dollar spent in the $0 to $2 range.

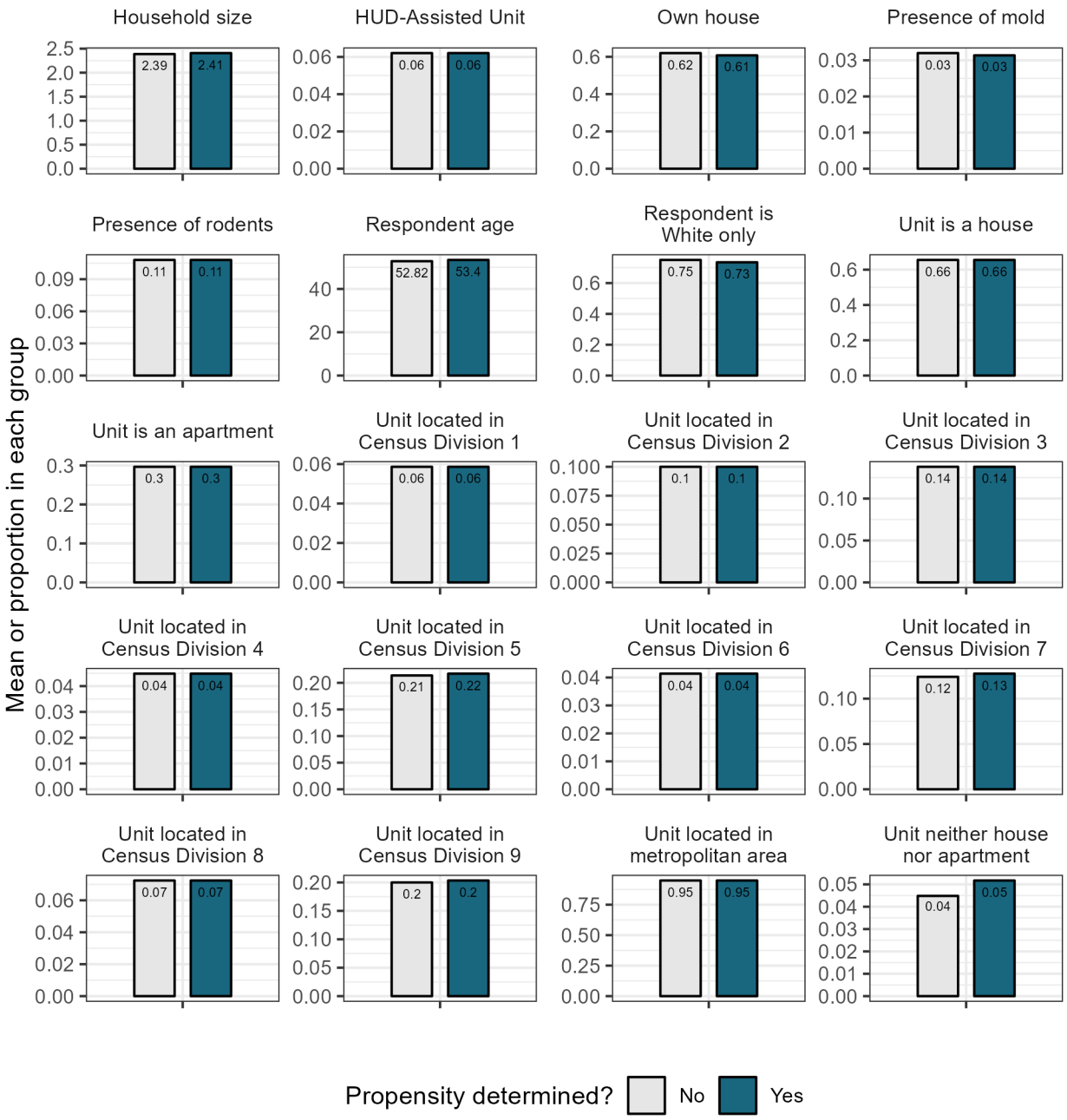
Findings on the Impact of Targeting Incentives

Finding 1: No Evidence That Targeting Incentives Reduces Nonresponse Bias Relative to Randomizing Incentives

Analysis bears no statistically significant evidence of differences in the estimated characteristics of the two randomly assigned groups: (1) those who received targeted incentive allocation and (2) those who received random incentives.

Figure 2 shows the examined attributes. Some small differences in attributes emerge—for instance, the sample produced because of targeting had fewer White respondents and more respondents from housing units that are neither houses nor apartments—but no clear differences in sample values.

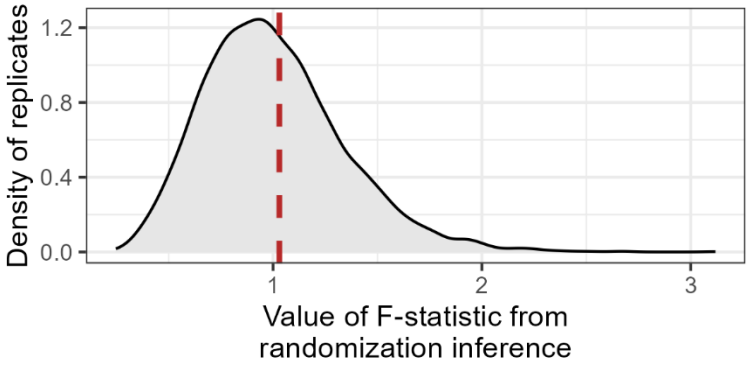
Figure 2. Attributes of Respondents from Targeted Allocation Versus Nontargeted Allocation



To reduce the chance of a false positive due to multiple testing, as preregistered, the null hypothesis that the difference in means between the two groups—the sample resulting from targeting and the propensity-independent sample—is zero across all the attributes that the figure depicts is tested using an omnibus test. Figure 3 shows the null distribution of F-test statistics. The observed test statistic is 1.03, and the p-value is $p = 0.42$, showing that an F-statistic as great as the one observed arises approximately 42 percent of the time due to the random assignment alone, even when the null hypothesis is true. This results in a failure to reject the null of no difference between the two conditions on all attributes. Overall, the evidence suggests that

the two groups—those incentivized using the targeted allocation and those incentivized completely at random—are similar in observed attributes.

Figure 3. Null Distribution of Test Statistics From F-Test From Randomization Inference Compared With Observed Test Statistic (Dashed Vertical Line)



Targeting did not appear to change *nonresponse bias,* at least along observable characteristics, because no statistically significant differences exist between the groups in the *sample composition*. The targeted incentives did not move the sample closer to or farther from population targets than the random incentives, because the two groups did not exhibit significant differences in their attributes. Therefore, it is unnecessary to use an external benchmark to investigate possible nonresponse bias changes.

Finding 2: Targeting Incentives Increased Response Rates Significantly More Than Randomizing Incentives but Did Not Reduce Effort

Although no statistically significant evidence showed that targeting incentives changed the sample average values of the assessed attributes, statistically significant evidence showed that targeting incentives increased the response rate. Importantly, each group received the same *quantity* of incentives. Therefore, this result comes from comparing the two strategies for allocating that same quantity of incentives.

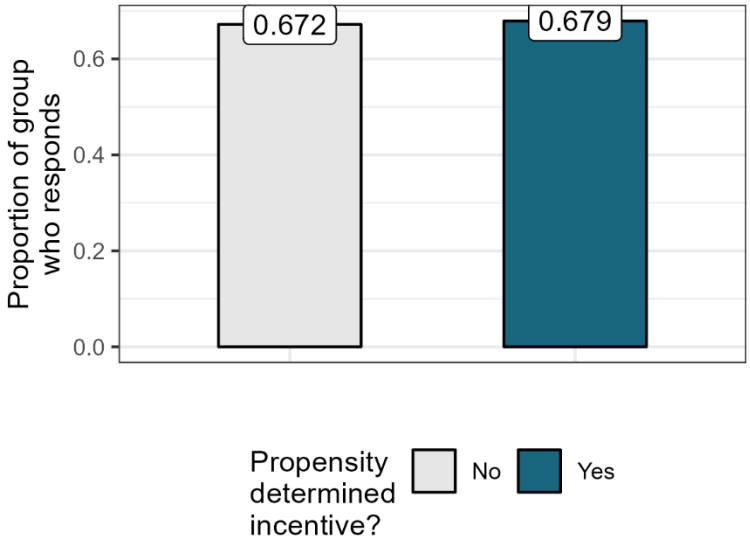
Figure 4 shows the raw differences in response rates. The estimates imply that targeting increased from a response rate of 67.2 percent in the nontargeted group to a rate of 67.9 percent in the targeted group, or 0.7 percentage point. The estimated difference is statistically significant at the $p < 0.05$ level (p = 0.0184 from randomization inference [RI]).[10] By translating this estimated difference to more concrete terms using the sample size of $N = 86{,}000$ respondents, targeting incentives using propensity scores induces an approximate additional 600 people to respond compared with randomly targeting the same amount of incentives.[11] In terms of variance, this increase in sample size implies a 0.7-percent reduction in sample proportion estimate variation, such as the proportion of individuals living in inadequate housing.[12]

[10] The parametric p-values also show that the estimated difference is significant at the $p < 0.05$ level.

[11] The sample size number is rounded according to U.S. Census Bureau entity counts guidance.

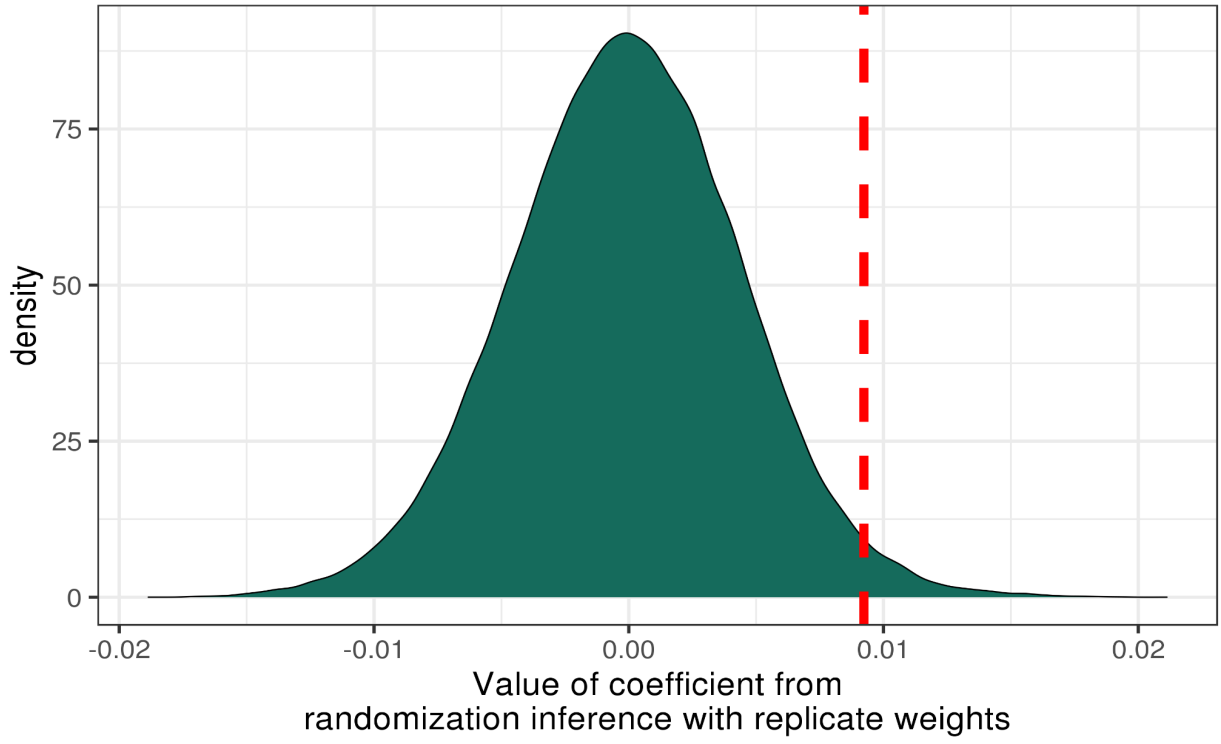
[12] The sample variance of a proportion with sample size n and probability p is $\frac{p(1-p)}{n}$. Using $p(1-p) = q$: $\frac{\frac{q}{86000} - \frac{q}{86600}}{\frac{q}{86000}} = 1 - \frac{q}{86600} * \frac{86000}{q} = 1 - \frac{86000}{86600} = 0.007$.

Figure 4. Response Rates for Targeted Incentives Versus Nontargeted Incentives



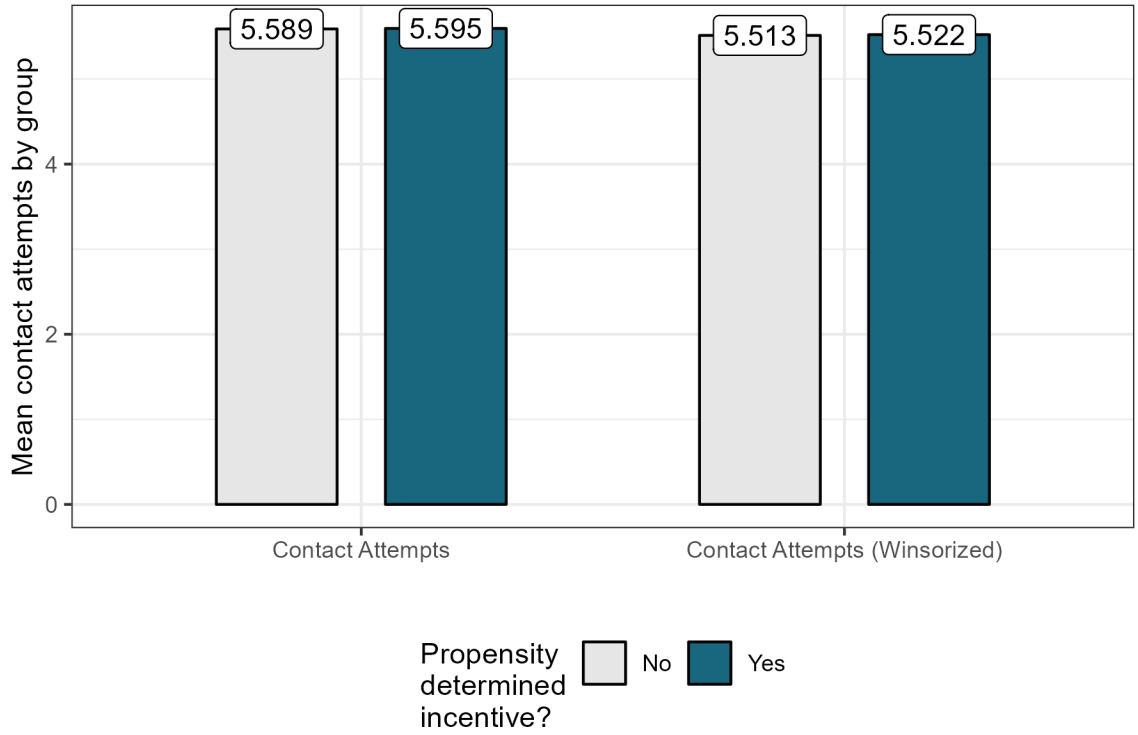
The results remain statistically significant when adjusted for not only the variance resulting from the experimental procedure *within* the American Housing Survey (AHS) housing units, but also for the variance from selecting *which* housing units are selected into the AHS by taking each of the 160 replicate weights, conducting RI with m = 1,000 rerandomizations for the estimates produced using each replicate weight and using the resulting 160,000-long list of estimates as the null distribution. The result remains statistically significant at the $p < 0.05$ level (p = 0.042 from RI), with Figure 5 showing the observed coefficient compared with the null distribution.

Figure 5. Null Distribution of Test Statistics From Regression of Response Rates on Incentives From Randomization Inference Compared With Observed Test Statistic (Dashed Vertical Line)



Although the evidence shows an increase in response rates, it is unclear how the relative response rates relate to the level of effort. It is possible that targeting incentives also changes the average number of contact attempts to achieve a response. Figure 6 shows two measures of the number of telephone and in-person contact attempts to arrive at those response rates: the raw number of contact attempts and a winsorized measure that codes contact attempts above the 99th percentile value to the 99th percentile value.[13] Neither of the estimated, between-group differences were statistically significant ($p = 0.7844$ for the nontransformed contact attempts measure; $p = 0.761$ for the winsorized measure). One interpretation is that the increase in response rates from targeting the incentives using propensities did not stem from extra effort on the part of survey enumerators. However, the propensity-targeted incentives also did not reduce the effort survey enumerators expended to induce a response relative to the random incentives. In other words, targeting the incentives slightly increased response rates for a statistically indistinguishable level of effort (and the same incentive amount). Simply changing *how* incentives were allocated appears to have increased the response rate and kept field costs constant.

Figure 6. Contact Attempts for Targeted Incentives Versus Nontargeted Incentives



Findings on the Impact of Different Incentive Amounts

Finding 3: No Evidence That Receiving Incentives Versus Not Receiving Incentives Reduces Nonresponse Bias

Turning to the impact of incentives on nonresponse bias, no statistically significant evidence showed that the characteristics of respondents who received incentives differ from those who did not. The value of the F-statistic is 0.882, which is statistically insignificant ($p = 0.8$). In other

[13] The 99th percentile value recorded 21 contact attempts.

words, no pattern of evidence suggested that incentives changed the sample composition—even without an external benchmark for population statistics with which to compare the sample estimates (such as mandatory surveys like the Decennial Census or American Community Survey). It follows that no evidence exists that incentives increased nor decreased nonresponse bias.

Finding 4: No Evidence That Incentives Increase Response Rates or Decrease Enumerator Effort Across the Sample as a Whole

The Finding 2 section showed that two different incentive allocation methods produced different response rates. Targeting incentives produced a response rate that was 0.7 percentage point higher than allocating an equivalent incentive budget completely at random. One implication is that incentives induced response—at least for some respondents in the group receiving the targeted allocation. What is less clear is the average effect of receiving an incentive across all respondents in the sample, irrespective of the allocation method. As explained previously (Treatment conditions and probability weights), the average effect of receiving an incentive on response rates and enumerator efforts in the entire sample is estimated by reweighting the data to account for different incentive receipt probabilities based on random allocation assignment methods. Note that some precision is lost because of the weights employed in this analysis.

Table 3 shows the estimated effects of incentives on response rates and contact attempts when comparing any incentives with no incentives, specific incentive amounts with no incentives, and the linear, marginal effect of an additional dollar across the incentive range—labeled approaches A, B, and C, respectively, in Table 3.

Table 3. Estimated Effects of Incentives on Response Rates and Contact Attempts

Approaches	Coefficient	Model	Estimated Effect	Standard Error	RI p-Value
A	Binary indicator for any incentive	Responded = binary indicator for any incentive	0.00485	0.004898	0.322
B	$2 (comparison 0)	Responded = factor form of incentive amount	– 0.001334	0.009355	0.9032
	$5 (comparison 0)	Responded = factor form of incentive amount	0.0004302	0.009342	0.9652
	$10 (comparison 0)	Responded = factor form of incentive amount	0.01135	0.006629	0.1146
C	Continuous incentive amount	Responded = continuous form of incentive amount	0.001053	0.0006483	0.135

Approaches	Coefficient	Model	Estimated Effect	Standard Error	RI p-Value
A	Binary indicator for any incentive	Total contact attempts = binary indicator for any incentive	– 0.06317	0.05037	0.2126
B	$2 (comparison 0)	Total contact attempts = factor form of incentive amount	– 0.09555	0.09259	0.355
	$5 (comparison 0)	Total contact attempts = factor form of incentive amount	0.0442	0.09355	0.668
	$10 (comparison 0)	Total contact attempts = factor form of incentive amount	– 0.1012	0.06745	0.1832
C	Continuous incentive amount	Total contact attempts = continuous form of incentive amount	– 0.007986	0.006596	0.2814

RI = randomization interference.

Approach A. The rows labeled "Responded = binary indicator for any incentive" and "Total contact attempts = binary indicator for any incentive" models report the results from the inverse propensity-weighted regressions of the response rate and contact attempts on a binary indicator that codes a respondent as "yes" if they received any incentive regardless of amount and "no" if they did not receive any incentives. The first row reports that receiving any incentive increases response rates by an estimated 0.5 percentage point, but this estimate is not statistically significant. Receiving any incentives leads similarly to approximately 1/20th of a contact attempt less enumerator effort, but this estimate is not statistically significant.

Approach B. Disaggregating the incentives into specific amounts, the rows describing the "Responded = factor form of incentive amount" and "Total contact attempts = factor form of incentive amount" models report the results of similar regressions of the response rate and contact attempts on binary indicators for A = 2, A = 5, and A = 10 (in addition to pair block indicators).

Receiving $2 versus $0 decreases the response rate by an estimated 0.1 percentage point and decreases enumerator effort by approximately 1/10th of a contact attempt. Neither estimate is statistically significant. Receiving $5 versus $0 increases the response rate by 0.04 percentage point and increases enumerator effort by approximately 1/20th of a contact attempt. Neither estimate is statistically significant. Receiving $10 versus $0 increases the response rate by an estimated 1 percentage point and decreases enumerator effort by approximately 1/10th of a contact attempt. Neither estimate is statistically significant ($p = 0.11$ in the randomization

inference). Taken as a whole, no statistically significant evidence shows the effectiveness of incentives in increasing the response rate.

Approach C. Finally, the remaining models regress the response rate and enumerator effort outcomes on a *continuous* coding of A. Each additional incentive dollar increases response rates equivalent to an estimated 0.1 percentage point and a marginal decrease in enumerator effort equivalent to less than 1/100th of a contact attempt. Neither estimate is statistically significant.

How should one reconcile the results reported in finding 2, where *targeting* incentives was estimated to produce a statistically significant 0.7-percentage point increase in the response rate with the statistically insignificant average incentive effect estimates in the sample as a whole? At least two nonrival explanations exist. First, it may be that incentives were only effective in the targeted group, because randomly assigning incentives implies sending them to respondents with a high likelihood of responding ex-ante. In this case, when the groups that received targeted and completely random incentives are pooled to analyze the average effect of incentives in the whole sample, the group that was allocated incentives completely at random drags down the average estimated incentive effectiveness. Second, the estimated average incentives effect may be less precise due to the reweighting required to estimate them. For example, the standard error of 0.005 in row A of table 3 is nearly as large as the estimated effect reported under finding 2 (0.007). In other words, the loss in precision from reweighting makes false negatives comparatively more likely, and the null findings reported here may reflect an insufficiently powered analysis.

Finding 5: No Evidence for Diminishing Marginal Returns to Incentives

The estimates presented in the previous section indicate that, overall, each dollar of additional incentives increases response rates and decreases contact attempts, although neither estimate is statistically significant. This section relaxes the assumption that each additional dollar produces the same marginal effect on response rates and effort. The coefficients A2, A5, and A10 estimated in the previous section are used to detect the presence of diminishing marginal returns.

Table 4 shows the results for the two preregistered comparisons.[14]

1. Comparing \$2 to \$5 with \$0 to \$2 by subtracting the marginal change in outcomes when moving from \$0 to \$2 from the marginal change in outcomes when moving from \$2 to \$5.
2. Comparing \$5 to \$10 with \$2 to \$5 by subtracting the marginal change in outcomes when moving from \$2 to \$5 from the marginal change in outcomes when moving from \$5 to \$10.

Table 4. Estimated Effects of Marginal Increases in Incentive Amounts on Response Rates and Contact Attempts

Outcome	Comparison	Difference in Marginal Return	p-Value
Response	2–5 versus 0–2	0.001255	0.8782
Response	5–10 versus 2–5	0.001596	0.7747
Contact attempts	2–5 versus 0–2	0.094360	0.2425

[14] As noted previously, the assumption of monotonicity implies that it is not necessary to estimate models comparing the change between \$0 and \$2 with \$5 and \$10.

Outcome	Comparison	Difference in Marginal Return	p-Value
Contact attempts	5–10 versus 2–5	– 0.075670	0.1899

Diminishing returns are signed differently depending on the outcome. A negative response rate value indicates that each dollar spent on incentives at higher levels produces a smaller increase in the response rate. A positive contact attempt value indicates that each additional dollar spent on incentives at higher levels produces a weaker reduction in enumerator effort.

Table 4 reports the results. Inconsistent with diminishing marginal returns, the first and second rows report positive values for the difference in marginal returns to the response rate. The estimates, taken at face value, imply that each additional dollar spent in the $2 to $5 range is 1/10th of a percentage point *more* effective than dollars spent in the $0 to $2 range. Each dollar spent in the $5 to $10 range is more effective by roughly the same margin. Neither result is statistically significant, however.

The third and fourth rows also provide no conclusive evidence for diminishing marginal returns to incentives with respect to enumerator effort. Taken at face value, the difference of 0.09, when comparing $2 to $5 with $0 to $2, implies that the marginal dollar spent in lower parts of the incentive range is more effective at reducing effort than in higher parts, but this relationship does not hold when considering the difference in marginal returns between $5 to $10 and $2 to $5. Neither estimate is statistically significant. The results suggest that the inflection point for incentives is greater than the $0 to $10 incentive amount range explored here if it exists.

Conclusion

This project represents an innovative attempt to evaluate two strategies for allocating the same quantity of incentives within the American Housing Survey (AHS): nonresponse risk-based targeted incentive allocation and completely random incentive allocation. The estimates suggest that targeting can significantly increase the response rate relative to random incentives, helping increase the sample size of respondents. However, corresponding changes in the sample composition do not accompany this improvement. The two samples—those incentivized to respond using targeted incentives and those incentivized to respond using random incentives—are similar in observed characteristics. The estimates suggest that targeting neither reduces nor increases bias in sample estimates relative to their population targets.

When interpreting the results, it is important to remember that features of the AHS panel may limit the generalizability of the results. Specifically, the targeted incentives theory assumes the ability to accurately target incentives based on characteristics of interest, for example, accurately predicting the most and least likely respondents and which respondents are the most likely to increase or reduce nonresponse bias. As opposed to a panel of people, a housing unit panel introduces additional noise to the predictive models due to the possibility of potential respondents within the housing units changing between AHS waves—approximately 28 percent of occupied households in 2019 had a member who moved during the past 2 years. Despite the use of rich data from prior AHS panel years, the nonresponse propensity scores used for targeting in this analysis had good—but not great—accuracy in predicting nonresponse (see the Overview of Experimental Design section). In other contexts, it is possible that different models—using different nonparametric models, different data sources, or different survey panel designs—could predict nonresponse better and be more useful for accurate incentive targeting. Exploring better models might provide fertile ground for future incentive research; however, the

results of this evaluation show that targeting incentives in the AHS, using the best available predictive model, was unsuccessful in reducing nonresponse bias relative to a random incentive allocation method.

References

U.S. Census Bureau and U.S. Department of Housing and Urban Development (HUD). 2020. "2019 AHS Integrated National Sample: Sample Design, Weighting, and Error Estimation." Washington, D.C., United States Department of Commerce, U.S. Census Bureau. https://www2.census.gov/programs-surveys/ahs/2019/2019 AHS National Sample Design, Weighting, and Error Estimation.pdf.

U.S. Department of Commerce and Department of Housing and Urban Development (HUD). 2015. AHS Integrated National Sample: Sample Design, Weighting, and Error Estimation. Washington, DC. https://www2.census.gov/programs-surveys/ahs/2015/2015%20AHS%20National%20Sample%20Design,%20Weighting,%20and%20Error%20Estimation.pdf.

Additional Reading

Hansen, Ben B, and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies," *Statistical Science* 23 (2): 219–236.

Jackson, Michael T., Cameron B. McPhee, and Paul J. Lavrakas. 2020. "Using Response Propensity Modeling to Allocate Noncontingent Incentives in an Address-based Sample: Evidence from a National Experiment," *Journal of Survey Statistics and Methodology* 8 (2): 385–411.

Appendix A: Nonresponse Bias Memo

Nonresponse Bias in the American Housing Survey 2015-2019

Prepared for

Office of Policy Development and Research

U.S. Department of Housing and Urban Development

Prepared by

Office of Evaluation Sciences

U.S. General Services Administration

Updated: August 21, 2020

Contents

Executive Summary

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 purpose of this memorandum is to explore whether and to what extent nonresponse bias is present in the 2015, 2017, and 2019 national AHS.

Evidence of Nonresponse Bias in the AHS

In Section 2, we present two independent sources of evidence for nonresponse bias in the AHS. First, national-level population estimates derived from the 2015 AHS diverge significantly from comparable population quantities measured by the 2010 Census. Even when employing weights designed to correct for nonresponse bias, the results suggest the 2015 AHS overestimates the share householders who own their house outright (no mortgage or loan), are white only, who are 65 or older, and who are members of smaller households.

Second, we present several forms of direct evidence illustrating that responding and nonresponding units have very different characteristics. Responders are ten percentage points more likely than responders to receive rental subsidies, for example, and are more likely to rent than to own. Whether taking attributes one by one or as a whole, the divergences between the measurable traits of responders and nonresponders are much greater than we would expect to see due to sampling variability alone.

Predicting Nonresponse and Refusal

In Section 3, we use a set of machine learning methods to examine how well characteristics measured for all units in the sample, taken from the sampling frame and the area in which they live, predict any form of nonresponse and refusal more specifically. The analyses yield three main insights. First, the models fare very well by conventional standards used to score machine learning prediction accuracy, bolstering our confidence in our ability to predict nonresponse. This is important for the design of incentive delivery mechanisms that target potential nonresponders.

Second, the results show that our models predict outcomes in 2017 better than in 2019 and that we are able to predict refusal better than nonresponse, more generally. Finally, the most important predictors are prior year response and levels of effort related to interviewing units (e.g., the number of contact attempts). Contextual features also help to predict nonresponse: it is more likely in areas with more frequent cold and cool days, for example.

Patterns of Partial Response

Section 4 goes beyond the binary distinction between response and nonresponse to look at why some questions are left unanswered by survey takers and why some units answer in one wave of the AHS panel but not others. Respondents are least likely to answer questions that appear sensitive or are otherwise difficult to answer without more information, such as those pertaining to the level of crime in the neighborhood. While questions posed later in the survey are more likely to go unanswered, we do not uncover strong evidence in support of the idea that this arises due to the additional time elapsed (e.g., due to interview fatigue).

Our analysis of which kinds of units respond in 2015 but dropout due to refusal in 2017 reveals systematic patterns using a rich set of data, since we are able to draw on the 2015 AHS responses. We find units with younger householders interviewed later in the 2015 survey were

most likely to drop out in 2017. A host of other characteristics measured in the 2015 survey are also associated with the probability of dropping out, but no clear pattern emerges.

Consequences of Nonresponse

Section 5 discusses some potential consequences of nonresponse bias for researchers using the AHS data. We show how panel attrition could affect estimates of important relationships, such as how income relates to housing adequacy. Among units that responded in 2015, those who would go on to respond in 2017 exhibit a very different relationship between income and adequacy than those who would drop out. Any analysis of longitudinal trends restricted to units who respond in both 2015 and 2017 would thus overestimate the negative relationship between income and adequacy, even when employing weights. Similarly, metropolitan-level estimates from the 2015 AHS differ from the 2010 Decennial Census in ways that matter more for some regions and for some variables than for others. Whereas those who own a house with a mortgage or loan owing are consistently undercounted in all metropolitan areas, the proportion of non-White householders is most severely undercounted in metropolitan areas located in the states of California, Arizona, and Texas. These results suggest that without a better understanding of nonresponse bias relative to their planned analysis (including choice of sample composition, variable selection, and level of geography), researchers may draw misleading conclusions.

* * *

The analyses included in the memorandum, taken as a whole, provide several data points to demonstrate evidence of nonresponse bias in the AHS. The analyses also show that nonresponse can be predicted, which suggests that interventions targeted at encouraging higher response rates among units likely to be underrepresented in the group of responders could help to reduce nonresponse bias.

Note: The results used in this memorandum were approved under Census Bureau Disclosure Review Board (DRB) approval numbers: CBDRB-FY20-373 and CBDRB-FY20-POP001-0179.

Nonresponse Bias in the AHS 2015-2019

Prepared by: Office of Evaluation Sciences,
U.S. General Services Administration

1 Introduction

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. The AHS provides important information on key features of the U.S. housing stock: how many people rent versus own their homes? How many are evicted? What proportion of units have adequate conditions, and what are the demographics of those who live in inadequate units?

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.[15] 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.[16]

Within the context of a panel survey like the AHS, nonresponse not only is declining, but also is a dynamic phenomenon. We refer to units where an interview does or does not take place as "responders" and "nonresponders" throughout this memo.[17] As Table 1 shows, of the 67,775 occupied units introduced into the national AHS sample in 2015, only 70 responded in both 2015 and 2017. Thirteen percent of those units in which someone was interviewed in 2015 were not interviewed in 2017, while 8 percent of those not interviewed in 2015 were interviewed in 2017. Another 8 percent of the occupied units sampled were never interviewed.

If the features we want to, but cannot, measure for nonresponders differ systematically from those of responders, nonresponse can lead to bias. If not addressed in some way, the presence of bias implies that the sample estimates will not converge to the true, underlying quantity in the population, no matter how large the sample of responders.

To account for this risk, the AHS calculates a nonresponse adjustment factor (NRAF) 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, should reduce or even remove the inferential threats posed by nonresponse

[15] See the OMB guidance at https://www.whitehouse.gov/wp-content/uploads/2021/04/standards_stat_surveys.pdf.
[16] The response rates for the 2015 and 2017 waves are taken from the AHS public methodology reports. The response rate for the 2019 wave is taken from our analysis of the IUF with the below restrictions to the national sample and excluding the bridge sample, with values based on the coding responders as STATUS == 1, 2, or 3 ($n = 63{,}186$) and nonresponders as STATUS == 4 ($n = 22{,}965$). These may differ from those in the published methodology report if there are different inclusion criteria for the published rates to remove ineligible households.
[17] In Section 1.1, we discuss distinctions between different types of units in each category—namely, within responders, occupied units interviewed at their usual residence versus vacant units versus units interviewed elsewhere. Similarly, nonresponders contain not only respondents who are found and who actively refuse, but also other categories. We discuss which types of responders and nonresponders we include in the different analyses.

bias. However, there is no guarantee that the model used for bias-adjusted estimates contains all the information it needs.

Table 1. Nonresponse Among Occupied Units Added to the Sample in 2015

Category	N Units
Interviewed 2015–2017	47,442 (70 percent)
Interviewed 2015, Not interviewed 2017	8,872 (13 percent)
Not interviewed 2015, Interviewed 2017	5,713 (8 percent)
Not interviewed 2015–2017	5,748 (8 percent)

The purpose of this memorandum is to understand whether and to what extent the 2015, 2017, and 2019 waves of the AHS exhibit nonresponse bias, with and without adjustment. Section 2 assesses the degree of bias both by comparing AHS estimates to the 2010 Census and by assessing whether the traits of responders differ from those of nonresponders. Section 3 delves more deeply into the sources of nonresponse, exploring how well we can predict nonresponse and which attributes of units and geographic areas are most predictive. In Section 4, we move from a binary measure—did the unit respond or not—to unpack the phenomenon of "partial" response: e.g., the fact that some units drop out of the panel in 2017 having had interviews in 2015, or the fact that respondents sometimes refuse to answer questions mid-survey. Finally, in Section 5 we give an overview of whether and how much nonresponse bias affects researchers' ability to estimate important relationships in the data and CBSA-level statistics.

1.1 A Note on Terminology and Method

This section briefly reviews some key terminology covering the different samples, interview types, and weights in the AHS, before providing an overview of how our analyses use different samples and weighting choices.

There are two broad categories of AHS samples: the national and the metropolitan sample. The national sample is a nationallyrepresentative biannual panel, whereas the metropolitan sample is comprised of a rotating series of large metropolitan areas. We focus on the national AHS sample.

In all analyses, we exclude the 6,000 units that are part of the break-in series or "bridge" sample, which are units that were part of the pre-2015 AHS panel left in to investigate the effect of changes to sampling introduced in 2015. We do not exclude other subsamples, such as the over sampling of HUD-assisted units, as these are accounted for in the weights employed throughout.[18]

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.

Noninterviews are split into three types: Type A noninterviews occur when a regular occupied interview or usual residence elsewhere interview fails, usually because the respondent refuses, is

[18] Put in terms of the AHS variable names, we exclude units with BRGSMPFLG == 1. We then include units if they *either* are part of the non-metro national sample (AHSCBSASUP == 6) or if they are part of the top 15 metros (AHSCBSASUP == 7 & TOP15FLG == 1).

temporarily absent, cannot be located, or presents other obstacles (such as language barriers the field staff are unable to overcome). Type B and Type C noninterviews both 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. We clarify below how these different categorizations are employed in the analyses.

Finally, we employ different kinds of weights in the different analyses. The AHS uses a four-stage weighting procedure. 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" 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).

The analyses in the present memorandum use two types of weights. For estimates that include only respondents, we employ the composite weight, WEIGHT, which is the final output of the above process, alongside the 160 corresponding replicate weights used to estimate the variance of sample statistics. We refer to this as the "composite weight" or "adjusted weight" throughout, as it adjusts not only for different probabilities of being sampled but also adjusts for potential nonresponse bias. In order to understand what nonresponse bias looks like when we do not try to explicitly adjust for it through the weighting scheme, we also employ what we refer to as the "base weight" throughout, which corresponds to the inverse sampling probability of each unit, or the first stage weight described above.[19]

Table 2 previews each of the analyses we report and the samples and weights used for each. In general, there were two forms of variation:

> **Which units are included in the analytic sample?** Analyses that rely on characterizing demographic features of responders focus on (1) responders who are (2) classified by the STATUS variable as an "Occupied interview," or as a responder who is not a vacant interview or usual residence elsewhere. Analyses that rely on sampling frame features generally focus on (1) responders regardless of their classification (including URE and vacant interviews) and (2) nonresponders regardless of their reason for nonresponse (including not only refusals but also nonresponses due to other codes). Finally, other analyses focus specifically on contrasting occupied interview responders with refusers.

[19] Put in terms of the AHS variables, the composite weight refers to the combination of the WEIGHT variable and the replicate weight variables REPWGT.*. The base weight refers to the BASEWGT variable.

Are the estimates reweighted and, if so, how? We describe whether and how we reweight observations using the two types of weights described above—the base weights that only account for differential probabilities of being sampled and the composite weights that account for both those differential probabilities of selection and nonresponse adjustment factors.

Table 2. Analyses, Samples, and Weights Used

Analysis	Which sample(s)?	Reweight?	Rationale
Evidence of Nonresponse Bias			
Benchmarking to Decennial Census (Section 2.1)	2015 AHS respondents (Occupied Interviews only)	Compares base weight to composite weight	2015 since most proximate to Decennial. Occupied Interviews for comparability.
Differences in Attributes (chi-squared; attribute by attribute) (Section 2.2)	2015, 2017, 2019 (analyzed separately); all respondents and nonrespondents	Compares unweighted to base weight	Examines sampling frame attributes relevant for all rather than demographic attributes less relevant for URE/vacant interviews
Differences in Attributes (R-indicator; summary measure across attributes) (Section 2.3)	2015, 2017, 2019 (analyzed separately); all respondents and nonrespondents	Base weight	Examines sampling frame attributes relevant for all rather than demographic attributes less relevant for URE/vacant interviews
Predicting Nonresponse and Refusal			
Predicting nonresponse (Section 3)	2017 wave; 2019 wave; all types of nonresponse and interviews	None (Section 3 discusses)	General nonresponse
Predicting refusal (Section 3)	2017 wave; 2019 wave; refusals and occupied interviews only	None	Refusal as specific behavior
Patterns of Partial Response			
Item order and partial completion (Section 4.1)	2019 wave; responders only	None	
Partial completion via attriting from panels (Section 4.2)	2015 is focal wave; 2017 refusal; focus on occupied interviews and refusals	Composite weight	Responders only
Consequences of Nonresponse			
Attritor heterogeneity analysis (Section 5.1)	2015 is focal wave; 2017 attrition; focus on occupied interviews and refusals	Composite weight	Responders only
Metro-level benchmarking (Section 5.2)	2015 wave; responders only	Composite weight	Responders only

1.2 Informing Experiments to Reduce Nonresponse Bias

A second goal of this memorandum, in addition to characterizing nonresponse bias in the AHS, is to explore possible predictors of and mechanisms for nonresponse bias. Understanding the predictors of nonresponse bias is useful for informing interventions to reduce nonresponse bias. Specifically, this memo informs an intervention designed to target incentives at units most likely to contribute to nonresponse bias with the goal of differentially increasing responses among those units to achieve more accurate survey estimates. As we discuss in greater detail later, one important consideration in targeting any intervention is whether the unit (or more precisely a person who resides within) is likely to be a "never responder"—that is, they never respond even if targeted with an intervention—or has characteristics that indicate amenability to interviews given the right approach. This might suggest modeling specific forms of nonresponse, such as refusal or attrition between panels, if we think these forms of nonresponse are more susceptible to intervention.

2 Evidence of Nonresponse Bias in the AHS

2.1 Comparing 2015 AHS Sample Estimates to the 2010 Census: National-Level Analysis

Background

A simple way to test whether the characteristics of a sample diverge systematically from the population from which it is drawn is to compare the population-level estimates with known population-level quantities. Here, we leverage the fact that the American Housing Survey and 2010 Census provide nationally representative statistics on adult householders to understand whether and to what extend the AHS sample estimates diverge from 2010 Census counts.[20]

The 2010 Census defines a "householder" in the following manner:

> One person in each household is designated as the householder. In most cases, this is the person, or one of the people, in whose name the home is owned, being bought, or rented and who is listed on line one of the questionnaire. If there is no such person in the household, any adult household member 15 years old and over could be designated as the householder.

The AHS definition of a "householder" parallels that used by the 2010 Census:

> The householder is the first household member listed on the questionnaire who is an owner or renter of the sample unit and is 15 years or older. An owner is a person whose name is on the deed, mortgage, or contract to purchase. A renter is a person whose name is on the lease. If there is no lease, a renter is a person responsible for paying the rent. If no one meets the full criteria, the age requirement is relaxed to 14 years or older before the owner/renter requirement. Where the respondent is one of several unrelated people who all could meet the criteria, the first listed eligible person is the householder. In cases where both an owner and renter are present, the owner would get precedence for being the householder.

We focus on how well national estimates of householder characteristics from the 2015 AHS align with 2010 Census summaries of the same characteristics. Statistically significant

[20] Note that the person who responds to the AHS survey and provides demographic information about the householder may not necessarily be the householder.

differences may arise due to nonresponse bias, but also through subtle differences in the definitions or methods used to identify householders, or due to demographic changes during the five-year period between the 2010 Census and the 2015 AHS. This analysis therefore provides an exploratory assessment of how much nonresponse bias may exist in national-level estimates but does not conclusively establish that such differences are due to nonresponse bias.

Methods

To calculate national estimates from the AHS, we first subset the 2015 internal use file to nonvacant interviews[21] that are not part of the bridge or metropolitan samples.

We take two approaches to weighting national average estimates: the first weights responses only by the inverse of the probability the unit was sampled; the second weights responses by the composite weight used to account for differential nonresponse in the AHS (see section 1.1 above). Comparison of the estimates derived from the two weighting schemes is informative about how well the nonresponse adjustment factors and raking schemes account for possible nonresponse bias.

To estimate the variance of the sample mean estimates, we employ the standard replicate weights contained in the internal use file.[22] For each feature of interest, this procedure provides a weighted mean estimate from the AHS and its standard error estimate. We treat the 2010 Census measure of the characteristic as a known population mean (e.g., with variance of zero) and derive a p-value through a one-sample, two-sided t-test of the null hypothesis that the sample mean is equal to the population mean.

Results

The results are reported on Figure 1. Points correspond to the difference between 2010 Census figures and sample-weighted (circles, `BASEWGT`) and bias-adjusted (triangles, `WEIGHT`**)** population mean estimates from the 2015 AHS, with positive numbers indicating possible overrepresentation. Numbers on vertical line centered at 0 correspond to 2010 Census means. Horizontal lines indicate 95 percent confidence intervals derived from standard errors estimated through BRR replicate weighting.[23] When these do not overlap the vertical line centered at 0, we interpret the difference to be statistically significant (i.e., highly unlikely to arise due to sampling variation alone).

The number centered at zero on the first row indicates that 19 percent of householders interviewed in the 2010 Census owned their house outright (without loan or mortgage). Taking the analysis at face value, the circular point on this row indicates that the 2015 AHS contains roughly eight percentage points "too many" such householders. Bias adjustment helps somewhat, bringing the sample estimate closer to the population statistic. However, even with bias-

[21] The Decennial Census focuses on "Occupied Housing units": "housing unit is classified as occupied if it is the usual place of residence of the individual or group of individuals living in it on Census Day, or if the occupants are only temporarily absent, such as away on vacation, in the hospital for a short stay, or on a business trip, and will be returning." In the AHS, this is equivalent to focusing on non-vacant, usual residence occupied interviews.
[22] Specifically, we use Fay's Balanced Repeated Replication (BRR) method with $\rho = .5$ as described in (Lewis, 2015). This involves using both the WEIGHT variable and the 160 replicate weights.
[23] For this analysis, OES had access to replicate weights corresponding to WEIGHT but not for BASEWGT and used the former to estimate the uncertainty for both kinds of point estimate. Thus, it is possible that the width of the confidence interval around the estimates derived using the BASEWGT sample weight would be different if the correct replicate weights were used.

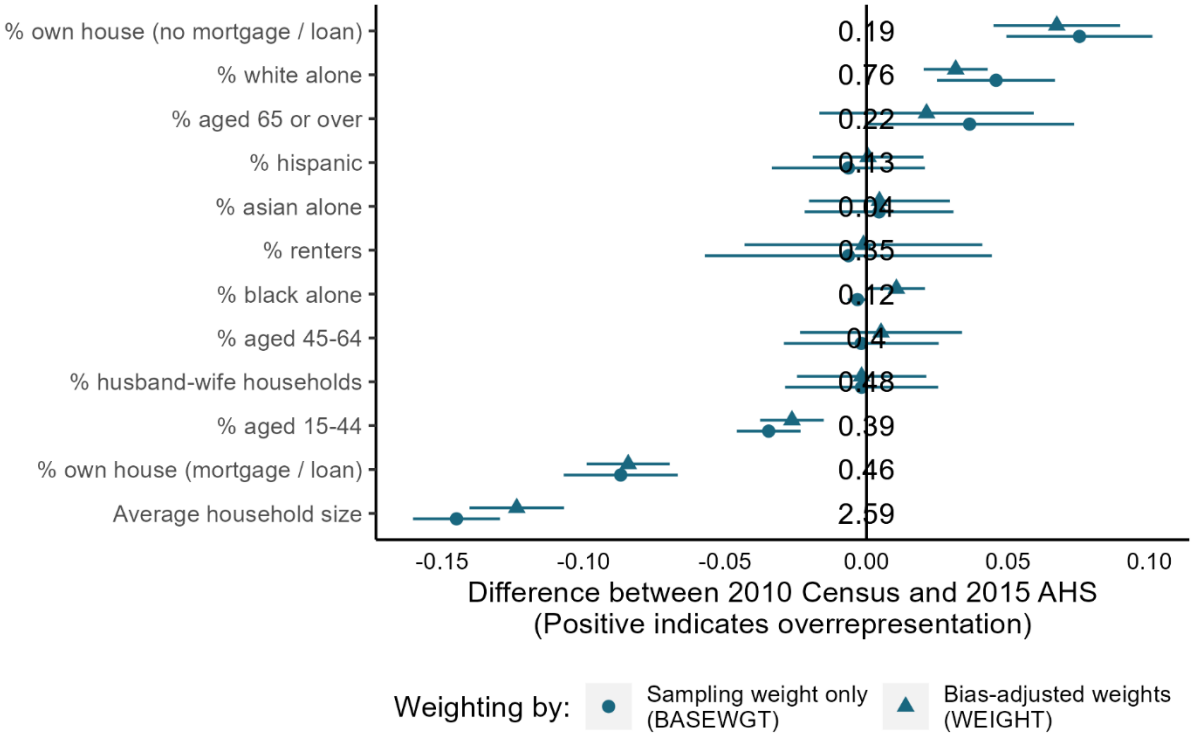
adjustment, the analysis presents statistically significant evidence that the estimate of the proportion of householders who own their property outright is too high.[24]

Similarly, the proportion of householders identified as "white alone" is five percentage points higher in the 2015 AHS than in the 2010 Census. Again, bias adjustment helps somewhat, but does not remove the discrepancy completely: whereas the Decennial Census indicates 76 percent of householders nationally are white alone, the bias-adjusted AHS estimate puts this number closer to 79 percent. Of course, it is possible that these divergences stem from demographic changes over time, so we should be careful in interpreting them as strong evidence of nonresponse bias. However, the direction of demographic change between 2010 and 2015—a lower national proportion of non-Hispanic white alone—could also mean we are underestimating the degree of bias.

[24] In feedback OES received on this finding, Census staff expressed the view that factors outside of nonresponse bias (differences in collection, sources of measurement error, imputation, etc.) could have been greater contributors to this difference than nonresponse bias. Relatedly, Census staff noted that other surveys have documented a downward trend in homeownership between 2010 and 2015. This would support the interpretation that the 2015 AHS overestimates the proportion of those who own houses outright. Further analyses could be done to address the limitations in this study: using more informative benchmarks with representative samples that are close to 100%, for example, or using more temporally proximate rounds of the Decennial Census and AHS (such as the 2020 Decennial and 2021 AHS).

Figure 1. Divergence Between the 2010 Census and National Estimates Derived From the 2015 America Housing Survey

Points correspond to the difference between 2010 Census figures and sample-weighted (circles) and bias-adjusted (triangles) population mean estimates from the 2015 AHS. Numbers on vertical line centered at 0 correspond to the 2010 Census. For example, first row indicates that 19 percent of householders interviewed in the 2010 Census own their house outright (without loan or mortgage), while bias-adjusted estimates from 2015 AHS estimate this proportion is roughly 7 percentage points larger (26 percent). Horizontal lines indicate 95 percent confidence intervals derived from standard errors estimated through BRR replicate weighting.



As we move down the plot from those two items, the attributes shown in the middle of the plot (e.g., percent Hispanic; percent husband-wife households) do not appear to diverge strongly from the 2010 Census. The weights produce a notable effect on the proportion of black householders: the unadjusted divergence suggests the AHS slightly undercounts this group whereas the adjusted estimate overcorrects and suggests an overcount. Finally, the results suggest that people aged 15 to 44 and large households are underrepresented in the AHS sample.[25]

2.2 Chi-Square Tests of Differences Between Responders and Nonresponders

Background

The previous section suggests that 2015 AHS estimates of population characteristics diverge significantly from counts in the 2010 Census. Divergences like this can arise due to nonresponse bias, but also due to actual demographic changes between the 2010 Census and 2015 AHS or the methodology used to sample householders. To assess whether nonresponse itself may play a role,

[25] In feedback provided to OES, Census staff clarified that the household size is expected to be an underestimate in the AHS, because the estimates are adjusted back to the housing-unit level in the final rake.

we can investigate whether units that respond to the survey are systematically different from those that do not. This section looks at which attributes differ between the two groups.

Methods

As Table 2 notes, the analytic sample is (1) comprised of all responders and nonresponders (regardless of whether the response was an occupied interview, URE interview, or vacant interview and the reason for the nonresponse), (2) includes each of the three waves, with the analysis conducted separately for each wave. "Unweighted" refers to estimates without any form of reweighting. The purpose of these estimates is to show how the differences in attributes in the raw sample will tend to get smaller as weights are applied to adjust for certain forms of oversampling. "Weighted" refers to estimates reweighting only by the inverse probability of selection (`BASEWGT`).

Since all the sampling frame variables we examined are categorical, we use a Chi-square test to test the null hypothesis that the frequencies of responders and nonresponders within each of the attribute levels is randomly and independently distributed. If the p-value indicates that the observed Chi-square statistic is highly unlikely given this null hypothesis (e.g., less than 5 percent), we interpret this as statistically significant evidence that the focal attribute is not independent of response status.[26] Statistically significant evidence of divergences between responders and nonresponders constitutes suggestive evidence of nonresponse bias, insofar as these characteristics are correlated with other important measures in the AHS.

The graphs show the following differences in proportions:

Proportion of responders (r) that fall into a given category of some attribute (l): $\frac{N_{lr}}{N_r}$, (e.g., the proportion of responders who fall into the "New England" category of the geographic division attribute).

Proportion of nonresponders (n) with that level of attribute (l): $\frac{N_{ln}}{N_n}$, (e.g., the proportion of nonresponders who fall into the "New England" category of the geographic division attribute).

Difference between #1 and #2: a positive point estimate indicates that the attribute level is overrepresented among responders.

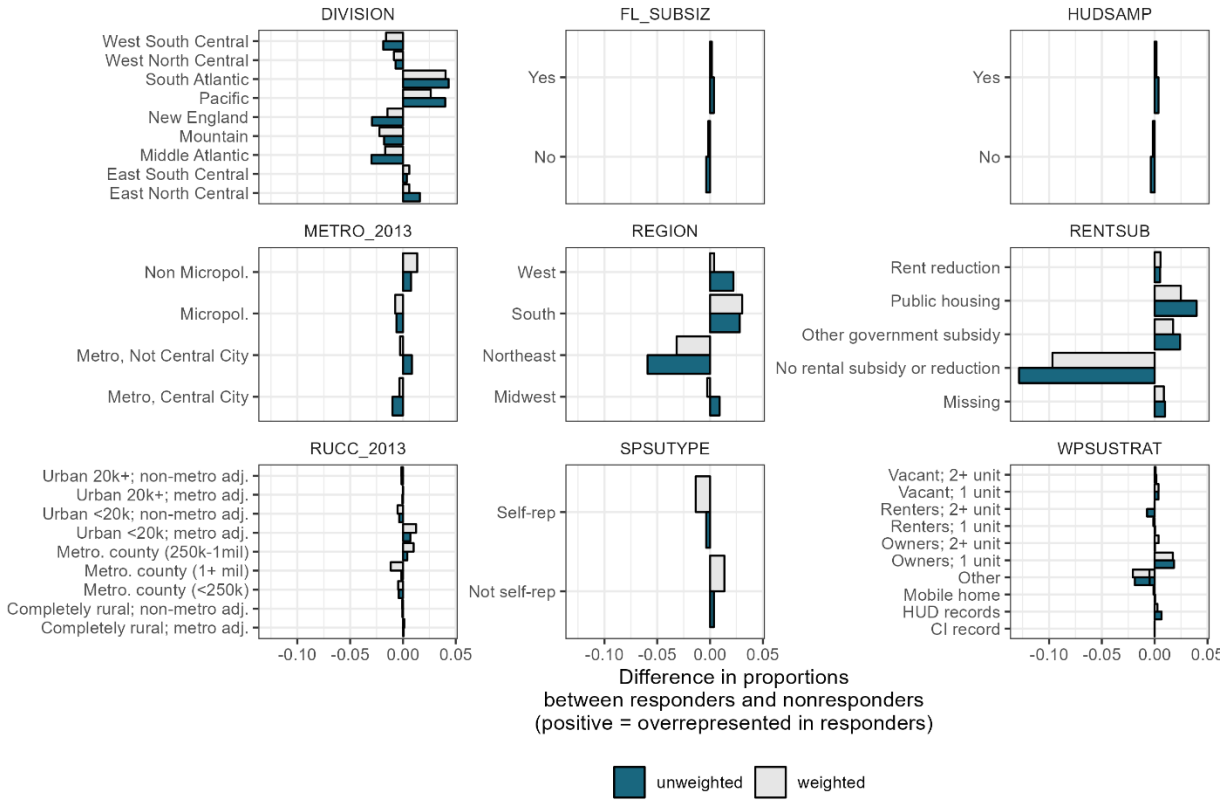
Results

Figure 2, which focuses on the 2019 wave, shows how responders differ from nonresponders for several sampling frame attributes. The dark bars represent the point estimates without weighting; the light bars represent the point estimates reweighting for unequal probabilities of selection (but without any noninterview adjustment factors applied). The figure shows that the probability of selection reweighting makes responders and nonresponders look much more similar by Census division, presence of a rental subsidy, and region. However, the graph also highlights the difficulty of balancing along many attributes. For instance, the reweighted estimates show more imbalance among self-representing versus non self-representing units than the unweighted ones.

[26] Note that this test does not take account of the clustering and stratification involved in the sampling design and makes an anticonservative assumption of independent sampling. OES was unable to take account of survey design due to the unavailability of replicate weights for nonresponders. This may be a useful area for further analysis and replication.

Finally, even after this initial reweighting, the two groups still look significantly different.[27] Results for the 2015 and 2017 waves look substantively similar, and all differences were statistically significant at the $p < 0.001$ level.

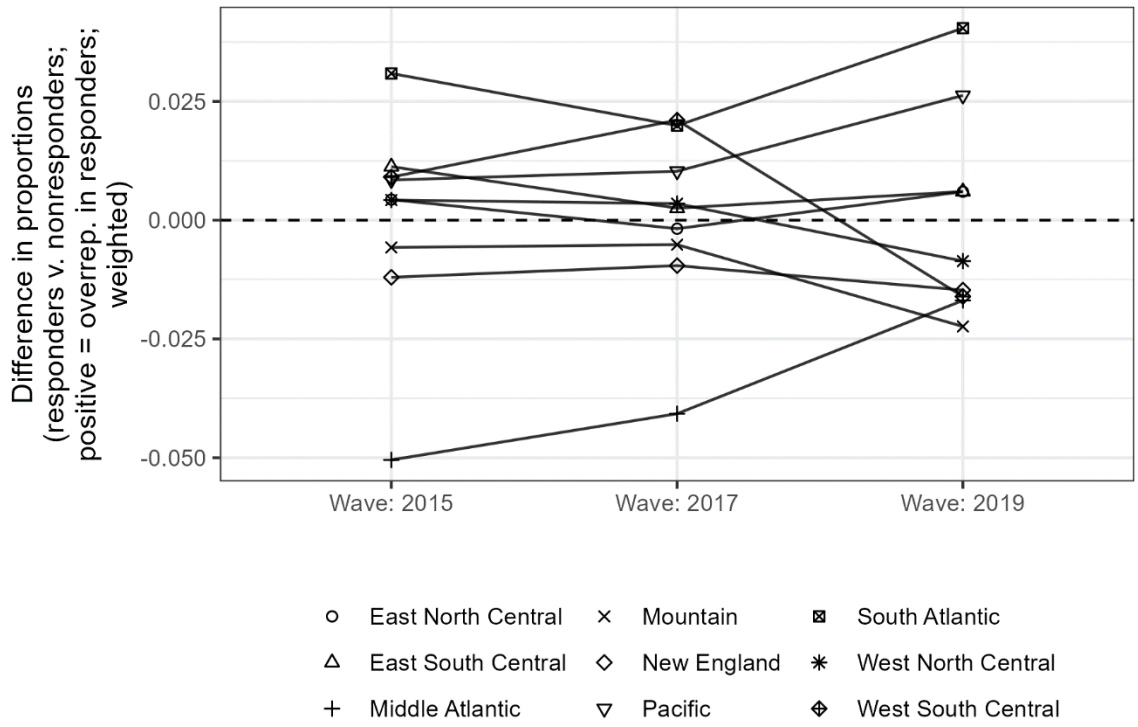
Figure 2. Differences Between Responders and Nonresponders: 2019 Wave
The figure shows the extent to which a level of an attribute is overrepresented in responders relative to nonresponders. Results for the 2015 and 2017 waves are similar and are found in Appendix Section A.1.



Finally, Figure 3, focuses on the Census division in which the unit is located and uses the weighted proportions to examine whether the differences vary across waves. We focus on Census division because of its importance in later-stage adjustments for nonresponse bias. The figure shows that while regions tended to stay on the same side of the red line indicating equal representation—that is, they tended to either consistently be over (above the line) or under (below the line) represented among respondents—some regions stayed fairly consistent in having similar proportions of responders and nonresponders and other regions fluctuated more—for instance, the Middle Atlantic region moving closer to equal representation. This analysis suggests not only that nonresponse bias likely present but also that it is dynamic and can shift in magnitude and possibly direction over time.

[27] Appendix Table 8 shows the p-values for each of the tests.

Figure 3. Changes Over Time in Over- Versus Underrepresentation
The figure focuses on the Census division variable and shows variation across waves in the extent of under versus overrepresentation.



2.3 Representativity Analysis

Background

In addition to the attribute-by-attribute analysis presented in the previous section, we can estimate an overall measure of how the observed attributes of responders differ from those of nonresponders. Schouten, Cobben, Bethlehem, et al. (2009) propose such a measure they call the "R-indicator." At its base, the R-indicator provides a standardized summary measure of whether observable characteristics of responders differ systematically from those of nonresponders.

Methods

The R-indicator is calculated as follows:

1. Estimate a binary regression predicting "interviewed" or not, based on attributes observed for both respondents and nonresponders (S),
2. Using the regression parameters from Step 1, predict each unit's propensity to respond, $\hat{y}$,
3. Find the standard deviation of predicted response propensities, $SD(\hat{y})$,
4. To get a value between 0 and 1, reparametrize so that:

$$\hat{R} = 1 - 2 \times SD(\hat{y}).$$

Provided we have good measures of the attributes of people who do not answer the survey, higher values of $\hat{R}$ indicate responders and nonresponders are similar, lower values indicate they are dissimilar. This approach relies on the availability of good measures observed for both kinds

of units, such as area-level characteristics or administrative data from other sources. It also relies on a well-specified model to relate the observed attributes to response status.

To understand the intuition behind the measure, consider the following thought exercise. Suppose there is a response rate of 50 percent, but the model is unable to detect anything systematically different about the responders and nonresponders. In this case, the prediction for each unit in the sample will be the same: $\hat{y} = 0.5$. As such, $SD(\hat{y}) = 0$, which implies $\hat{R} = 1 - 2 \times 0 = 1$. When $\hat{R} = 1$, our model is telling us whether or not someone responds is as good as random, so those who respond provide a good representation of those who do not, even with a 50 percent response rate.

Suppose instead that we were to discover that everyone who answered the survey had a first name starting with J, and none of the nonresponders had a first name starting with J. If we include an indicator for having a first name starting with J in our model, it will perfectly predict response: Jill, Jamal, and Julia, for example, would be predicted to respond with probability $\hat{y} = 1$, while Robin, Shaun, and Sara would have probability $\hat{y} = 0$, implying $SD(\hat{y}) \approx 0.5$, thus $\hat{R} = 1 - 2 \times 0.5 = 0$. So, conditional on having the right predictor for nonresponse, $\hat{R}$ tells us how well responders represent nonresponders. Note that $\hat{R}$ does not tell us how well responders and nonresponders represent the target *population*, only if the two groups are similar.

Of course, perfect prediction hardly ever happens in practice: just by random chance, we might end up with a large amount of people whose name starts with J who happen to respond, even if there is no true underlying correlation between these phenomena. Given the possibility that random sampling can produce meaningless correlations, the question is whether the correlations we observe in our model are greater than we would expect to observe just by chance alone. Values of $\hat{R}$ that are really unlikely to occur just due to random chance, say less than 5 percent, are "statistically significant."

To infer the probability of getting the $\hat{R}$ we observe, we need to estimate the variance of $\hat{R}$. Schouten, Cobben, Bethlehem, et al. (2009) derive the standard error of $\hat{R}$ through resample bootstrapping. In order to obtain confidence intervals, they assume that $\hat{R}$ is normally distributed. However, our analyses suggest these standard errors are not amenable to the typical Z-score transformation used to obtain p-values in T-tests.

We therefore use a permutation test in order to make an inference about whether we would expect to see the observed $\hat{R}$ simply by chance, or whether the observed $\hat{R}$ is statistically significant. Specifically, we randomly shuffle the variable indicating response and reestimate the $\hat{R}$ hundreds of times in order to obtain some of the $\hat{R}$ values we might have estimated if there were truly no correlation at all between the predictors and the outcome. We compare this distribution to the observed $\hat{R}$ to get a p-value corresponding to a one-sided test: the probability of observing just by chance an $\hat{R}$ at least as low as the one we observed, supposing that there is no true relationship between nonresponse and our predictors. We calculate this probability by taking the proportion of permuted R-indicators at least as low as the observed one.

We consider 73 predictors that are available for both responders and nonresponders sampled into the 2015, 2017, and 2019 AHS surveys. These include variables from the sample frame, such as the whether the housing is HUD-assisted, as well as information about the Census tract level in which the potential respondent is located drawn from the American Community Survey (ACS),

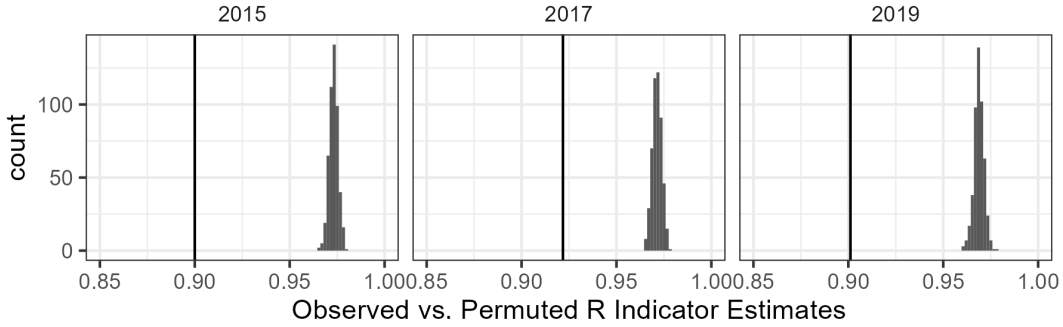
such as the proportion of the units that are rented or the proportion of people who are white.[28] These ACS features are important because we know little about the demographics of "never responders." Note that we are measuring characteristics of areas rather than characteristics of a particular household in that area.

Results

Figure 4 plots the observed value for the R-indicator (thin vertical line) alongside the distribution of permuted R-indicator estimates (histogram). For all waves under analysis, the estimated R-indicator is below one and much lower than we would expect to see just due to random chance. In other words, we find statistically significant evidence that responders and nonresponders differ on a host of observable characteristics. Table 11 in Appendix Section 4.3 reports the numerical results: across the three waves, the R-indicator ranges from 0.90 to 0.92.

Figure 4. Evidence of Systematic Differences Between Responders and Nonresponders Across a Range of Predictors

Thin vertical line indicates estimated R-indicator. Gray histogram represents distribution of estimated R-indicators under the null hypothesis that response is independent from all predictors. The results are "statistically significant" insofar as the observed R-indicator is highly unlikely to arise due to chance alone under the null of independence.



One concern is that the permutation procedure does not faithfully describe the sampling variation, which might produce misleading p-values. Since the R-indicator analysis is simply another way of answering the question "how well does the model predict the data," we can also use a more conventional approach to hypothesis testing called a Likelihood Ratio Test (LRT). In essence, this test asks whether adding the predictors to an intercept-only model improves predictions more than we would expect by random chance alone. The results, also presented on Table 11 in the appendix, confirm the main finding: statistically significant evidence that nonresponders' attributes differ from those of responders.[29]

2.4 Section Summary

This section presents strongly suggestive evidence of nonresponse bias in the AHS. The 2015 AHS national estimates depart from corresponding population-level counts in the 2010 Census in key areas such as householder race and ownership status. Of course, divergences such as this may arise for reasons unrelated to the systematic exclusion of certain groups from the sample.

[28] Some predictors had to be dropped due to collinearity, which arises when two or more variables contain very similar or equivalent information on units in the analysis, and thereby "cancel out" each other's estimated influence on the outcome.

[29] As with other analyses involving nonresponders, these estimates of variance do not include replicate weights. This is an area for further analysis and replication.

However, in an analysis of a host of attributes available for those who do and do not respond to the survey, such as their housing type and the demographic characteristics of their neighborhood, we find strong evidence that responders look different from nonresponders. Analyzed either one by one or taken as a whole, the attributes of responders systematically differ from those of nonresponders. Future analyses could explore how much of the gap remains when we adjust estimates with the nonresponse adjustment factor.

3 Predicting Nonresponse and Refusal

Background

The R-indicator analysis in the preceding section uses attributes available for both responders and nonresponders to predict where nonresponse is most likely to occur. It does so using a fairly limited predictive method: a parametric model where (1) attributes about units enter additively into the model and (2) the model does not perform variable selection, or regularization that "zeroes out" the influence of attributes that do a poor job of predicting nonresponse. Many better methods for predicting binary outcomes exist.

The goal of the present analysis is to use a series of more flexible classifiers for two purposes. First, we predict which units will be nonresponders or refusers in a given wave of the AHS. Second, we focus on the top-performing models to explore which features of units best predict nonresponse and refusal.

Methods

The analysis focuses on prediction of one of two binary outcomes.

General nonresponse:

1 = `nonresponder`: for any reason (Types A, B, and C).

0 = `responder`: this includes (1) occupied interviews, (2) vacant interviews, (3) URE interviews.

Refusal:[30]

1 = nonresponder: due to refusal (subset of Type A nonresponse).

0 = responder: occupied interview only. Since occupied interviews provide the most direct contrast with refusals, the analytic sample excludes nonresponders who are not refusers as well as vacant and URE interviews.

We fit a series of binary classifiers to predict these two outcomes.[31] Table 3 outlines the classifiers, which fall into two general categories.

First 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 (dt.* in Table 3)—imagine we start with two features: the Census region in which a unit is located and the percentage of households with a

[30] This outcome is similar to the one used in the panel attrition analysis discussed below in Section 4.2. It differs in that it includes "never responders," whereas the panel attrition analysis is subset to those who responded in the 2015 wave.

[31] We chose classifiers using useful list for data science applications: https://github.com/rayidghani/magicloops.

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. The n_estimators argument changes the number of trees in the forest.

Finally, we employ gradient-boosting models (gb.*) and adaptive boosting (ada). These are two *ensemble classifiers*—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.

The second category of classifiers are *regularization based*. We use different forms of the lasso procedure, which is designed to strike a middle ground between selecting too many and too few variables as the best predictors of nonresponse and refusal. The procedure employs "penalized" regression. Put simply, the algorithm tries to fit a model with a "good" score. As its predictions of nonresponse and refusal get more accurate by adding better predictors, its score improves. However, for each variable the algorithm adds to a model, the score decreases—there is a penalty for including more predictors. In theory, if the degree to which new predictors are penalized is calibrated correctly, the algorithm will include the minimal set of variables that do a good job of predicting the outcome, while excluding those that do not add to the predictive accuracy, either because they are redundant (collinear with already included variables) or do a poor job of predicting.[32]

[32] All of these classifiers were fit in Python 3.6 using scikit-learn.

Table 3. Models Used to Predict Nonresponse and Refusal in Full Sample

Shorthand	Longer Description	Parameters
Tree-based models		
dt_shallow	Shallow decision tree	DecisionTreeClassifier(random_state=0, max_depth = 5)
dt_deep	Deeper decision tree	DecisionTreeClassifier(random_state=0, max_depth = 50)
rf_few	Random forest with fewer trees	RandomForestClassifier(n_estimators = 100, max_depth = 20)
rf_many	Random forest with more trees	RandomForestClassifier(n_estimators = 1000, max_depth = 20)
gb_few	Gradient boosting with fewer trees	GradientBoostingClassifier(criterion= 'friedman_mse', n_estimators=100)
gb_many	Gradient boosting with many trees	GradientBoostingClassifier(criterion= 'friedman_mse', n_estimators=1000)
ada	AdaBoost	AdaBoostClassifier()
Regularization-based models		
logit	Logit	LogisticRegression()
logitcv	Logit with penalty term selected via cross-validation	LogisticRegressionCV()
logitl1	Logit with L1 penalty	LogisticRegression(penalty = "l1")

We fit these models to two sets of features.

1. AHS-only features from two sources:
   a. **AHS sampling frame or master file variables.** We use 48 binary indicators created from each categorical level of the following variables:
      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 9. 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.

Finally, we evaluate the models using 5-fold cross-validation. The sample is randomly split into five evenly sized groups. Then, the model is fit to the data obtained by pooling four of the five groups (the training set). That model is used to generate predictions in the fifth, held-out group (the testing set). We use a set of evaluation metrics described below to measure how much those predictions in the fifth group deviate from the actual values the model is trying to predict. The process is repeated, using each fold as the held-out fold and calculating the scores each time. The results are averaged across the five folds.

We look at 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.[33]

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

Precision: $\frac{\text{TotalTP}}{\text{TotalTP+TotalFP}}$ Among predictions of nonresponders, what proportion are actually nonresponders;

Recall: $\frac{\text{TotalTP}}{\text{TotalTP+TotalFN}}$ Among actual nonresponders, what proportion do we correctly predict to be nonresponders, as opposed to erroneously predicting that they are responders;

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

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

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 use the F1 Score as the main 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 3 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).[34] 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.

Table 4. 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.

[34] The choice of threshold can be calibrated to balance precision with recall. The results presented use Auto-Sklearn's threshold for each of the models, which is generally 0.5. Next steps might involve better calibrating the threshold to a value that corresponds to the number of units we can target in an incentive experiment targeting units likely to be underrepresented in the responses (e.g., the 10,000 units with the highest predicted probability of nonresponse).

ID	Pred. $\hat{y}$ continuous	Pred. $\hat{y}$ binary	True y	error_category
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.

3.1 How Well Can We Predict Nonresponse and Refusal?

Results

Figure 5 focuses on predicting general nonresponse in the 2019 wave and shows that we are able to predict nonresponse with a high degree of accuracy.[35] Both types of approaches—regularization with the penalty chosen via cross-validation (logitcv); and tree-based approaches—performed well. The one model that performed less well was the "deep" decision tree. It is possible that this classifier overfit to the data because it used a single tree without a high number of predictors. Appendix Figure 22 shows the results for the 2017 wave, where our ability to predict is substantially higher than in the 2019 wave (mean F1 score across models of 0.88 in the 2017 wave compared to a mean F1 score of 0.85 in the 2019 wave). As we discuss in the section summary, this could affect how well we think we are able to predict nonresponse in the 2021 wave that will be the target of the proposed incentives experiment.

Comparing the predictions from the two types of features—features from the AHS only (including lagged response-related outcomes); those features and ACS contextual features—the contextual features from the American Community Survey (1) improve predictions across all classifiers but (2) these improvements are small, with only small increases in the F1 Scores for the models with ACS features compared to the models without. For the second, as the next section shows, *when included*, these ACS features are important predictors. This is likely due to a combination of reasons. First, the most predictive features in all models were lagged response-related variables—this lessens the predictive power of *either* ACS contextual features or AHS sampling frame features like region.

Second, since the sampling frame variables are largely geography based, they may capture similar information as the ACS contextual features.

[35] After fitting, we ended up excluding two of the logistic-based models from evaluation—logitl1 and logit—because while they had F1 Scores in the 0.80-0.85 range, the penalty parameters zeroed out nearly all of the predictors.

Figure 5. Ability to Predict Nonresponse: 2019 Wave
The figure shows F1 scores for two types of feature sets: AHS-only (which includes both sampling frame variables and lagged response/contact attempt variables) and those plus the ACS contextual features.

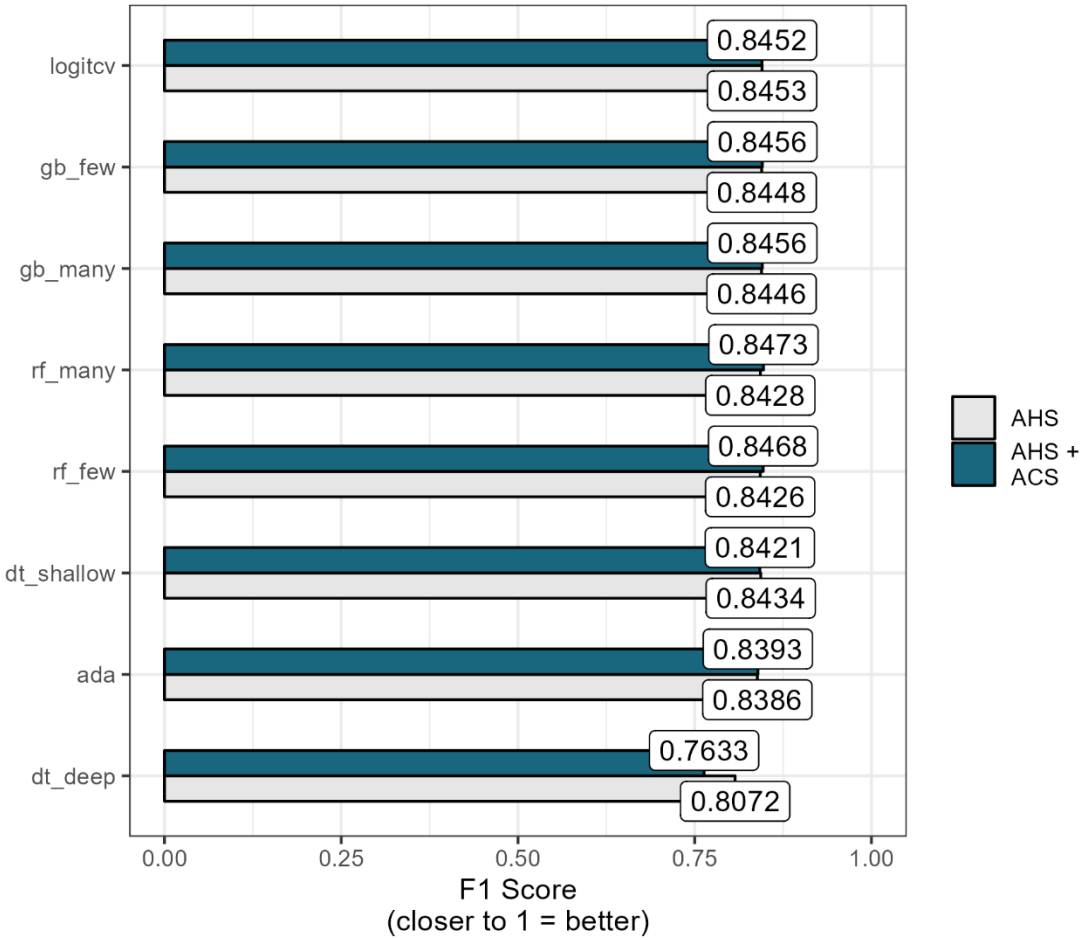


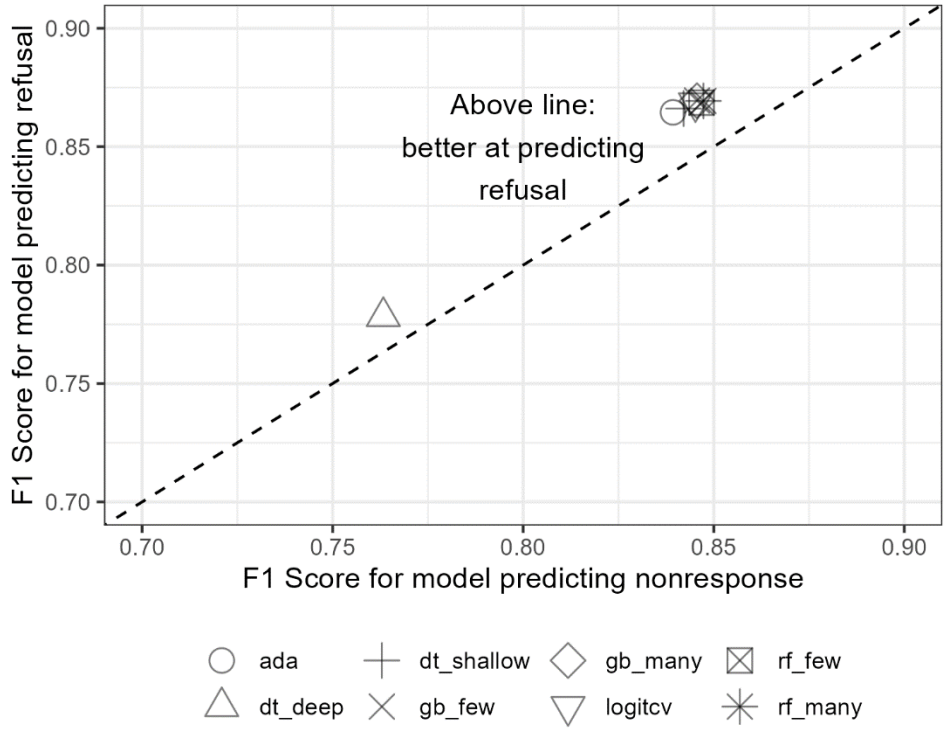
Figure 6 focuses on the contrast between our ability to predict nonresponse (previous graph) and our ability to predict refusal. Each dot represents one of the final models. The fact that all models are above the 45 degree line shows that while the predictions of nonresponse and refusal each have high F1 scores, the higher F1 scores of refusal indicate that we are better able to predict that outcome. Appendix Figure 23 shows the raw scores for 2017 and 2019 for the refusal models, for which we only estimated the models with combined AHS and ACS features. Similar to the results for nonresponse, the models show substantially better performance in the 2017 wave than in the 2019 wave.

As we explore further in the next section, our ability to predict refusal better than nonresponse could be driven by the fact that our most important predictor for nonresponse is whether the unit was a nonresponder in the previous wave; for refusal, our most important predictor is whether the unit is a refuser in the previous wave. In turn, since refusal is a narrower, more behaviorally rooted category than more general nonresponse,[36] we might be better able to leverage past

[36] As we discuss in Section 3.1, general nonresponse contains technical forms of nonresponse like Type C nonresponse (e.g., mobile home moved; permit abandoned) or other forms of Type A nonresponse like not being home.

refusal to predict refusal in a focal wave than past nonresponse to predict nonresponse in a focal wave.

Figure 6. Ability to Predict Refusal Versus Ability to Predict Nonresponse: 2019 Wave
Each dot represents a model. The x axis shows that model's performance in predicting nonresponse (relative to all types of response). The y axis shows that model's performance in predicting refusal (relative to occupied interviews). We see that the deep decision tree performs much worse than other models for each type of outcome. For all models, we are significantly better at predicting refusal than nonresponse.



3.2 Top Predictors of Nonresponse and Refusal

Results

The previous results show better, but not substantially better, performance when we include contextual features from the ACS. We can dig deeper into these patterns by focusing on two of the better performing models that yield different types of "top predictors": the random forest with many trees, which yields directionless feature importances, and the penalized logit, which yields more traditional coefficients that have a positive (predicts nonresponse or refusal) or negative (predicts response or nonrefusal) sign.

Importantly, and as in the panel attrition analysis we discuss later, all features are *predictive* rather than *causal*. For instance, there might be unobserved characteristics of a unit that lead that unit to refuse in both the 2015 and 2017 waves. The "did not respond in 2015" feature in the model predicting 2017 nonresponse is thus a proxy for those unobserved characteristics, rather than someone's nonresponse in a previous wave actively causing their nonresponse in a focal wave. Second, *within predictive features*, some yield more insight than others into mechanism for nonresponse or refusal. For instance, knowing that someone needed to be contacted five times before a response in the previous wave rather than just once may be highly predictive of

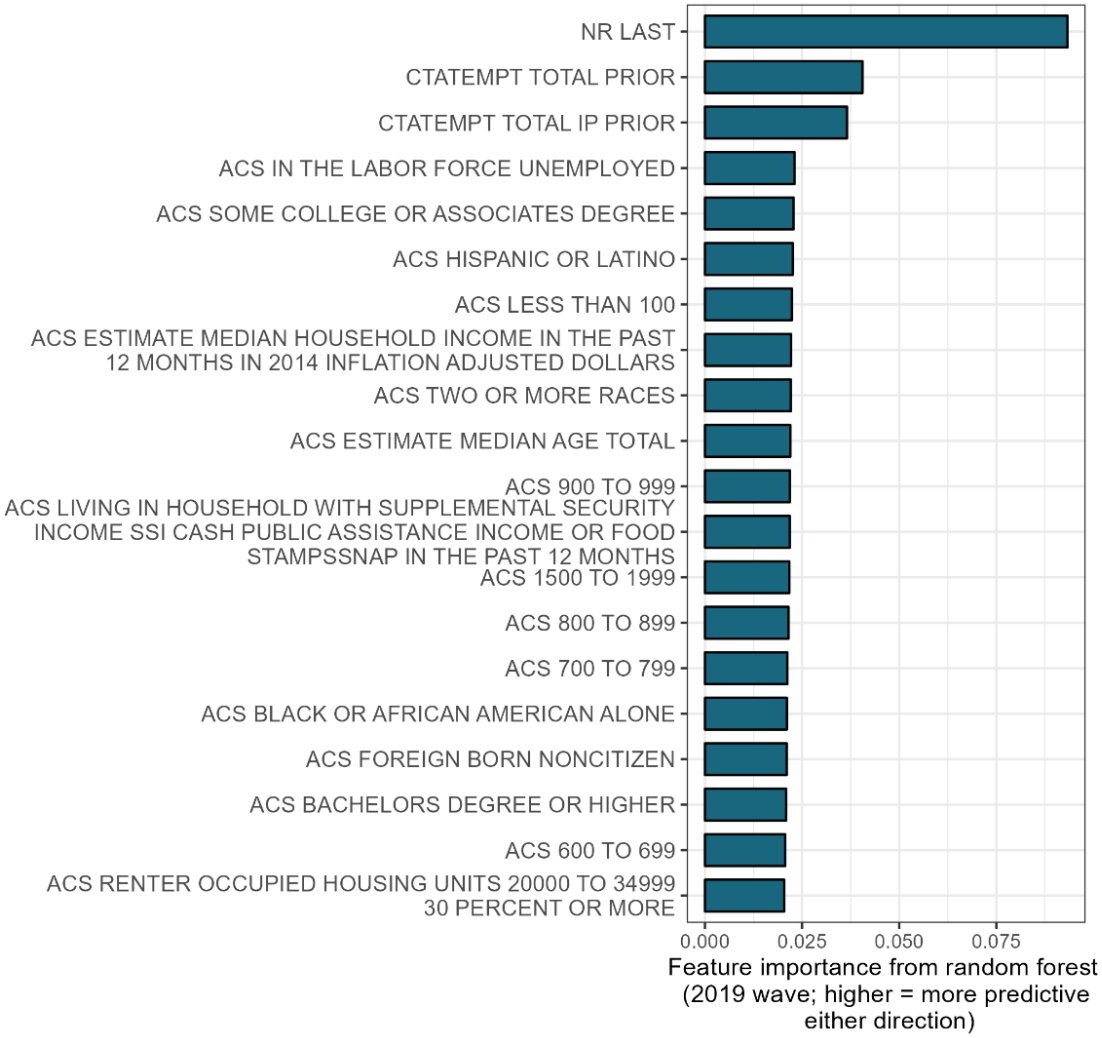
nonresponse in the focal wave. But we gain little insight into *why* they were both "reluctant responders" in the previous wave and nonresponders in the focal wave. In contrast, features like the ACS variables on the educational attainment of the local area, though possibly subject to ecological fallacy issues, could indicate more informative patterns.[37] In other words, it is more informative to know that area-level educational attainment is predictive of nonresponse because we may hypothesize that it relates to the level of trust in a government-sponsored survey, which can be addressed in an intervention, than to know only that a unit did not respond without additional information.[38]

Figure 7 shows the attributes with the top 20 feature importances for predicting *nonresponse* in the 2019 wave in the models using the combined AHS and ACS features, with Appendix Figure 15 showing the ranking of the remaining features. The figure shows that, perhaps unsurprisingly, the most important predictor of nonresponse in 2019 is 2017 nonresponse. Similarly, regardless of response status, the number of overall contact attempts and in-person contact attempts is highly predictive. These predictors fall into the category of useful if our goal is pure prediction, but they are arguably less informative for understanding mechanisms behind nonresponse. Contextual ACS features are perhaps more useful in generating hypotheses to explore. The local area's unemployment rate, age distribution, and monthly housing costs are all highly predictive. Yet, two limitations in interpretation remain. First, the the graphs reflects highly predictive features without direction—so, for instance, higher median age is highly predictive but we do not know from the model alone whether it predicts response or nonresponse.

[37] The ecological fallacy occurs when we use aggregate data—in this case, data about Census tract characteristics—to infer things about individuals that are part of that aggregate. In the present case, we observe a general correlation between an area having higher educational attainment and that area having a lower likelihood of nonresponse. However, it could be the case that within areas with higher educational attainment, lower educational attainment individuals are actually the most likely to respond.

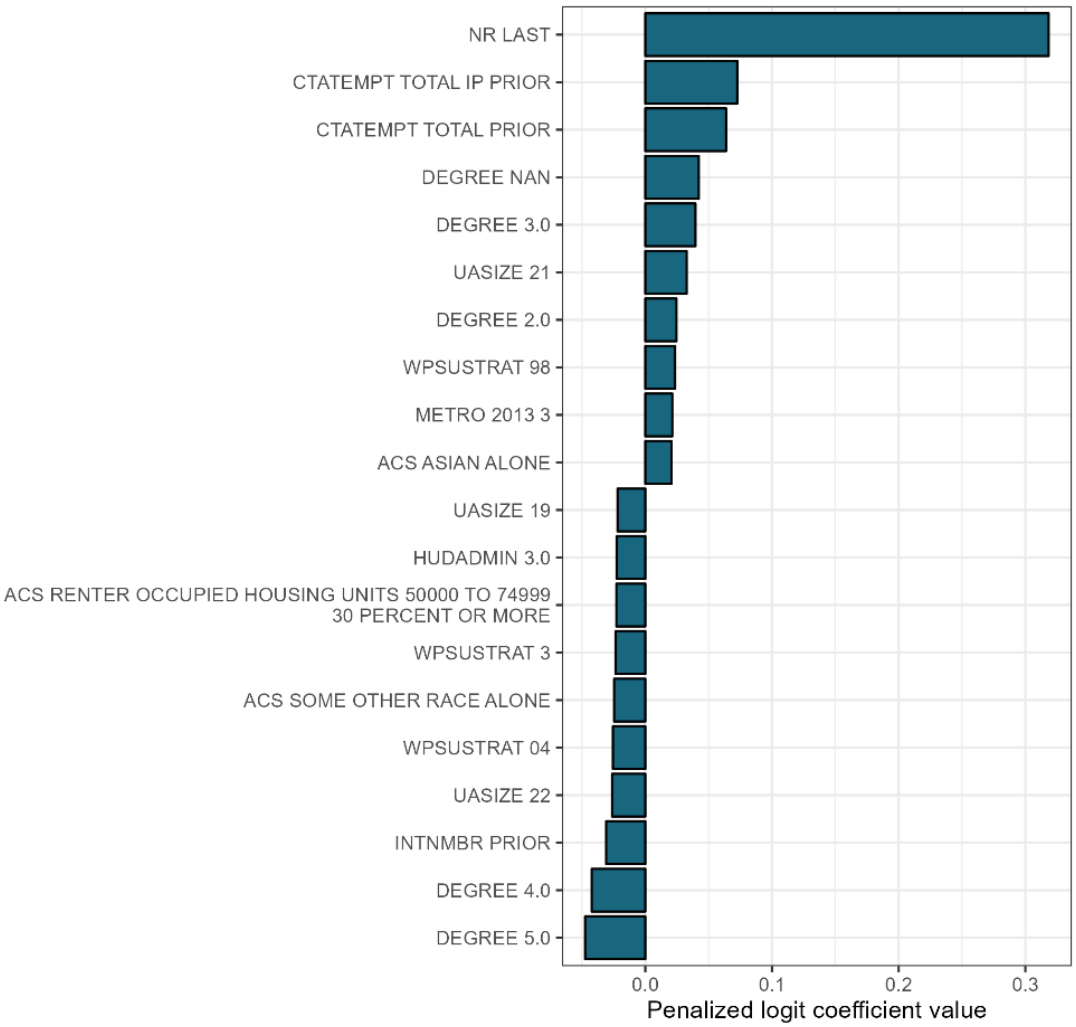
[38] The pattern—area-level lower SES is associated with a higher likelihood of unit-level nonresponse (Maitland et al., 2017)—has been observed in other social surveys. While trust is one mechanism, there might be many others like work schedules, time pressures, and more.

Figure 7. Most Important Predictors of Nonresponse—Random Forest Model: 2019 Wave
The figure illustrates the top 20 features. The ACS less than 100, 800 to 899, 1500 to 1999 variables refer to the dollar amounts of monthly housing costs.



To address the shortcomings of nondirectional feature importance, we turn to the top predictors from the penalized logit, which provides signed coefficients and has significant overlap in top predictors with the random forest model. Figure 8 shows the top 10 most highly positive (predictive of nonresponse) and highly negative (predictive of response) features from the penalized logistic regression. The results show that most of the highly predictive features in the random forest were highly predictive of *nonresponse* in the penalized logit—for instance, total contact attempts and prior nonresponse in 2017 are highly associated with 2019 nonresponse. In addition, features like DEGREE, which captures area-level temperature, show that areas with more cold and cool days have a higher likelihood of nonresponse (2 and 3, which represent areas where people need to use heat for a higher proportion of the year), and areas with mild or mixed temperatures have a higher likelihood of response. While these predictors may reflect patterns like the ease of in-person enumerators reaching households, they could also be proxies for unobserved characteristics of areas. Meanwhile, some features associated with a higher response level, like having more interview attempts in the prior wave, likely also reflect that units that respond in previous wave are likely to be responders again in the next wave.

Figure 8. Most Important Predictors of Nonresponse—Penalized Logit: 2019 Wave
The figure illustrates the top 10 positive and top 10 negative features.

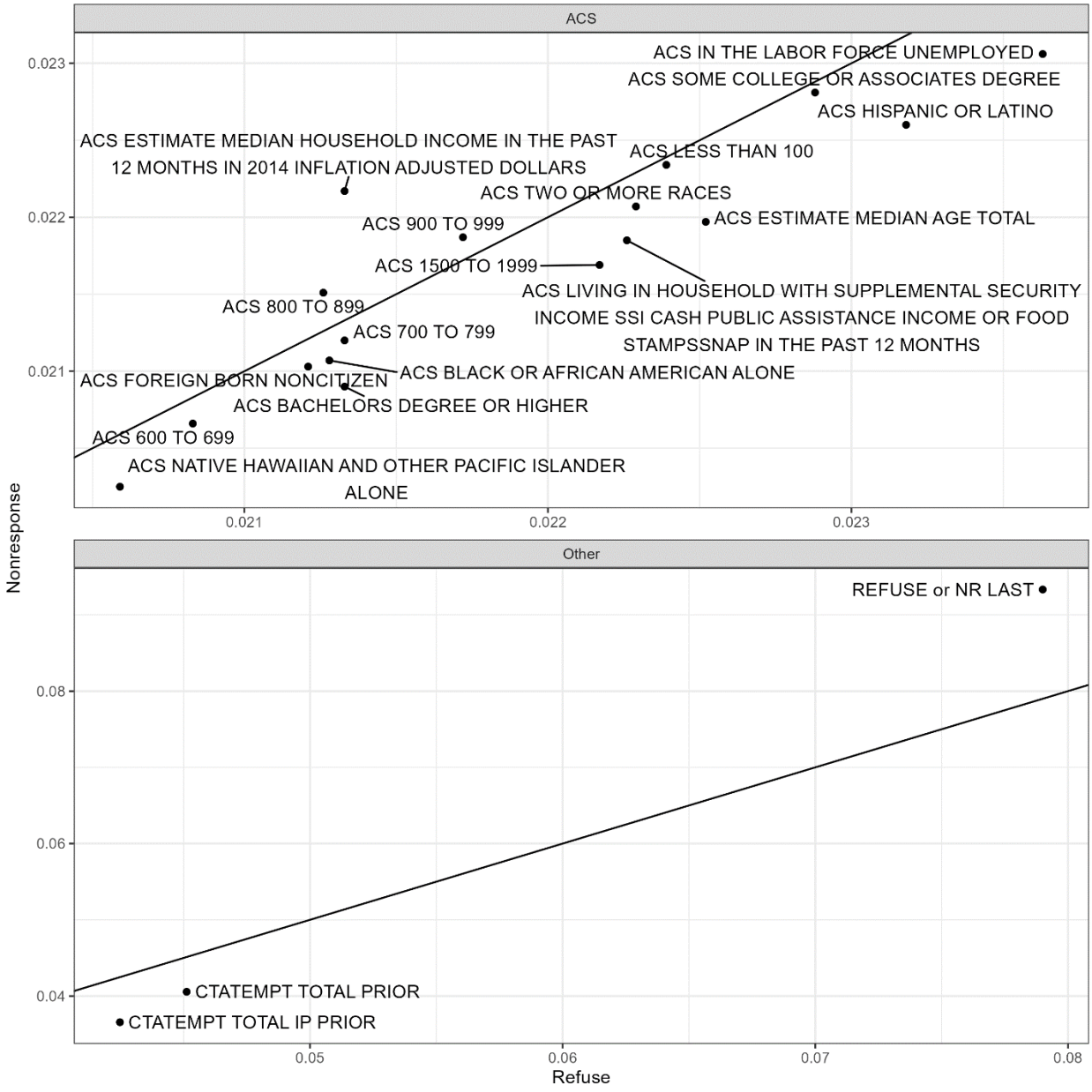


The second caveat in interpreting the results from the main models is that the model focuses on all forms of nonresponse. But as the previous section showed, we are better able to predict refusal than nonsreponse more generally. We next turn to the feature importances from the predicting refusal models to see which attributes remain important and which do not.

Figure 9 shows a scatterplot where each dot is a top feature from the random forest model predicting refusal. The x axis reflects its importance in the refusal model; the y axis its importance in the nonresponse model. Features above the 45 degree line are more predictive of nonresponse than refusal; features below of refusal. We see some patterns like the number of contact attempts in the previous wave being more predictive of refusal than of nonresponse. However, the generally high correlation shows that refusal and nonresponse are generally predicted by similar factors.

Figure 9. Top Predictors of Refusal Versus Top Predictors of Nonresponse—Random Forest: 2019 Wave

The figure illustrates the top 10 positive and top 10 negative features.



3.3 Section Summary

The results from the predictive models yield three main findings. First, we are better at predicting both nonresponse and refusal in the 2017 wave than in the 2019 wave. Second, we are better at predicting refusal than at predicting general nonresponse. As it relates to the planned intervention discussed in Section 1.2, the better ability to predict refusal could suggest a way to try to improve the efficacy of targeting and thereby reduce nonresponse bias. Namely, refusal is a behavior that we can potentially modify, but other forms of nonresponse may stem from nonbehavioral factors that are less likely to be affected by an intervention. For the intervention, we will consider carefully how the outcome we aim to predict—whether refusal or a more specific form of refusal like refusal over the phone but "yes" in person—correspond to different study goals. Third, the most important predictors of both nonresponse and refusal are the relatively "black-box" factors of a unit's status in the previous wave and its contact attempt history. These are arguably less useful for understanding mechanisms of nonresponse than some of the lower ranked ACS contextual features.

As we approach the proposed incentives experiment, we plan to dig more deeply into why our ability to predict is higher in 2017 than in 2019. One source could be the higher rates of nonresponse in 2019 than in 2017, which could reflect that the nonresponse and refusal categories contain a more heterogeneous mix of units. Another source is that we did not leverage the full panel nature of the data when constructing the "prior waves" variables. In particular, for the 2019 wave, our models only used the response status and contact history information from the 2017 wave; a better approach would be to construct features based on both the 2015 and 2017 waves. For 2021, we will have three waves of prior data and would be able to leverage the richer history for better prediction.

Second, prior to the experiment, we will delve more deeply into the unit-level predictions that generate the overall accuracy measures. For instance, which units are consistently flagged by all classifiers as having a high risk of nonresponse or refusal, versus which units' predictions are less stable across classifiers? How do the accuracy metrics vary by region? Questions like these can help pave the way for analytic decisions in the proposed experiment like whether to use a single classifier or whether, for instance, to use classifiers for different regions that perform well in those regions.

Finally, due to the focus on prediction and analytic challenges with including weights in the classifiers' estimation procedures,[39] the present results do not reweight the data. We may want to weight the data so that observations weighted more heavily via the AHS' weighting procedure are also weighted more heavily in the loss functions for each model.

4 Patterns of Partial Response

Beyond binary classifications of units as "responders" and "nonresponders" in a given wave of the AHS, we can also classify units according to how their response status changes over time, either between waves or within the survey itself. The present section focuses on two forms of "partial response." First, our analysis of item-level missingness explores why, *within* a survey wave, some households complete enough of the survey to count as a responder but fail to

[39]In particular, sklearn classifiers, vary in whether they accept a sample_weights argument, making it more straightforward to first estimate a range of classifiers and then choose the top-performing one that also accepts survey weights.

complete many questions on the survey (Section 4.1). Second, our analysis of panel attrition analyzes why, *between waves*, units which respond one year drop out in subsequent waves (Section 4.2).

4.1 Characterizing Item-Level Missingness: Item's Content Versus Item's Order

Background

The AHS uses two methods to treat missing values:

1. The majority of variables for which there is item-level missingness have values imputed, with an ancillary variable then created, the "imputation flag" variable, that indicates which responders have imputed values for the respective variable. The main variable then contains these imputed values.
2. A smaller subset of variables is not imputed, and the main variable contains missing values.

Figure 10 shows the top 20 items with the most imputation.[40] Figure 11 shows the top 20 items, among those not imputed, that have the highest rate of nonreport. Focusing on items with high rates of nonreport, we see some patterns like potentially sensitive items about neighborhood safety or financial challenges.[41] For instance, the following high missingness items might be more sensitive:

1. `NHQPCRIME`: Agree or Disagree: This neighborhood has a lot of petty crime.
2. `NHQSCRIME`: Agree or Disagree: This neighborhood has a lot of serious crime.
3. `NUMMEMRY`: Number of persons living in this unit who have difficulty concentrating or remembering.

[40] This was calculated by (1) looking at variables that have the J prefix indicating an edit flag and (2) looking at the proportion of responses in the 2019 IUF file for responders that have a value of 2 for that edit flag variable.
[41] We also see others like interview mode and language that might be not reported for more survey administration-based reasons.

Figure 10. Top 20 Items With Most Missingness, as Indicated by Edit Flag Variables
The figure illustrates the top 20 items from the 2019 IUF with the highest rate of editing using the above criteria. We exclude items that are edited for over 50 percent of responders on the grounds that these items likely reflect constructed variables that reflect imputation due to that construction process rather than imputation due to respondents not answering. Some variables—like YRBUILT AND YRBUILT_IUF—represent the public use version of the variable and the IUF version (in this case, yrbuilt is a more aggregated categorical value of yrbuilt_iuf for disclosure reasons).

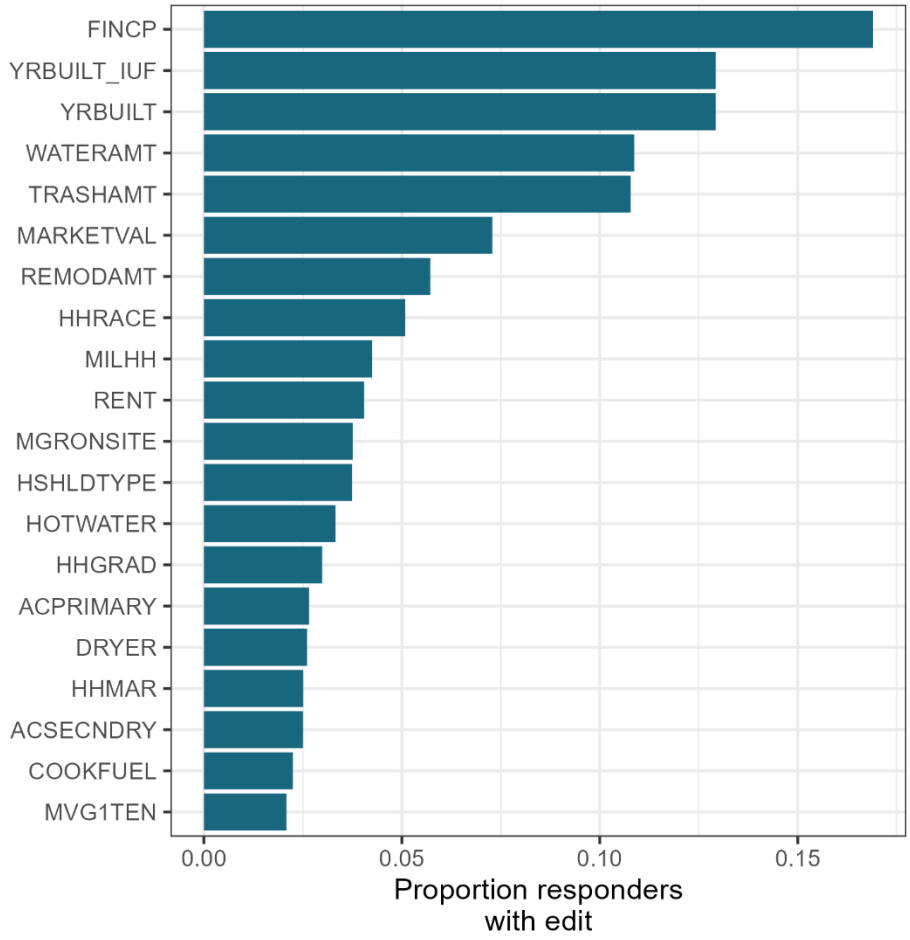
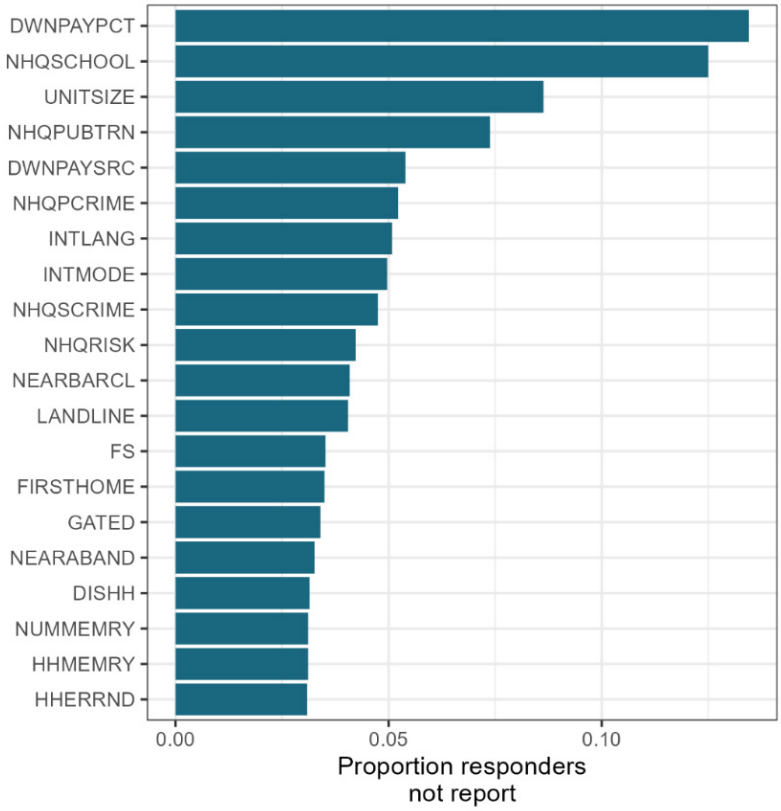


Figure 11. Top 20 Items With Most Missingness, as Indicated by Not Reported Values on Actual Variables

The figure illustrates the top 20 items from the 2019 IUF with the highest rate of "Not reported."



Yet, while "sensitive" questions might have higher rates of responders choosing to not report, there may be confounding at work. Namely, if the survey is designed to place less essential or more sensitive questions at the end, and if survey takers also get more fatigued and inclined to skip as the survey progresses, the correlations between item content and item missingness might be confounded by item placement. Put differently, among those who complete enough of the survey to count as responders, this missingness could stem from two sources:

- Missingness due to the item itself—for instance, sensitive questions having higher missingness.
- Missingness due to the item appearing later in the survey, a point at which respondents may have more survey fatigue and may either be (1) more likely to stop the survey altogether, or (2) complete the survey but skip more items to reduce time.

To examine these two possible sources of nonreport, we conduct an analysis of the impact of an item's order on nonresponse for that item. This analysis focuses on all items for which we can match the raw survey instrument names to the final analysis names, which includes some of the items discussed above as well as others we can match.

Methods

For these analyses, we use the 2019 trace file data. Each unit sampled has a text-based trace file that records the enumerator's keystrokes as they contact the respondent and move through the survey items. We parsed the trace files to extract the following information:

- The unit's identifier.
- The "instrument item name."

As we discuss below, the instrument name for items is sometimes distinct from the name the item is later given in the survey. Sometimes, there is a 1:1 mapping between an instrument item and survey item. Other times, multiple instrument items are combined to create a single survey item.

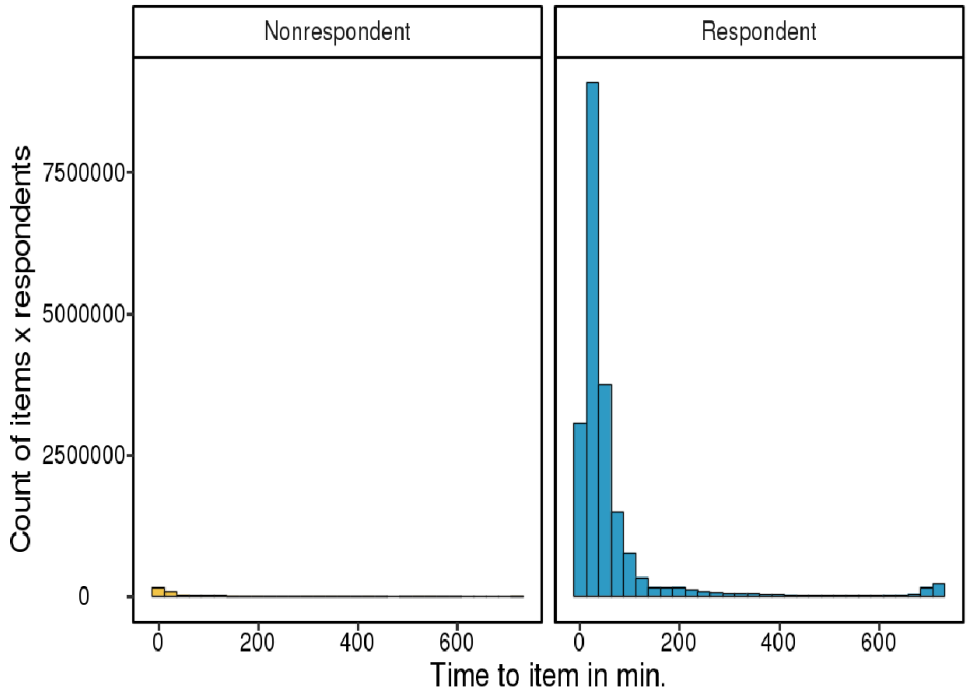
The parsing process was not perfect. In particular, for 18 units (less than 0.001 percent of the total occupied interviews included in the analysis), there were issues in how the timestamps were recorded that led us to exclude them from the analysis.

After parsing the files, we then created what we call the raw item duration. Broadly, this is the distance in time (minutes + seconds) between the focal item and the earliest timestamp for a particular day for that respondent (respondents can have interviews on multiple days if they start and stop the survey). More precisely, raw item duration is defined as follows, where i indexes a respondent, k indexes a particular item, and d indexes a calendar day:

$$\text{Raw item duration} = \text{timestamp}_{\text{idk}} - \min(\text{timestamp}_{\text{id}})$$

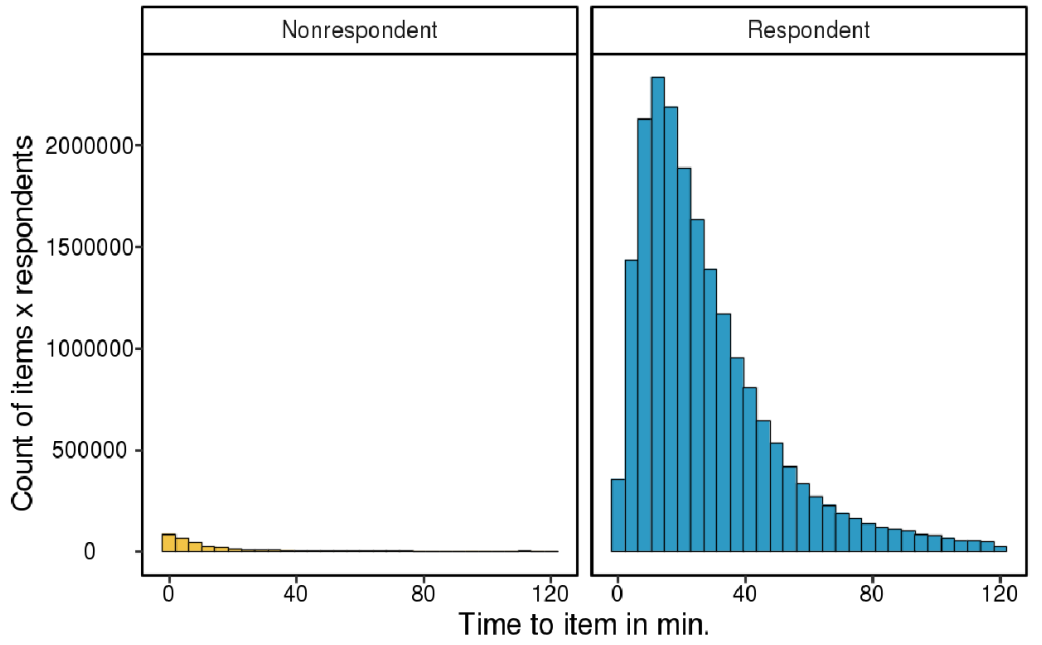
Figure 12 shows the distribution of durations using this raw measure. We see a bimodal distribution that stems from the fact that certain respondents have multiple interview sessions on the same day. This fact complicates measuring an item's duration from the day-specific minimum.

Figure 12. Distribution of Item-Specific Relative Durations (Raw)
The second small peak at closer to 8–10 hours shows that some respondents had multiple distinct sessions in the same day.



Due to this challenge, we use a rough measure of the start of the survey—the keystroke indicating the initiation of a new survey.[42] We then code what we call a cleaned item duration where a focal item is matched to (1) the nearest start of survey keystroke, that (2) is two hours or less away from that focal item. So if a respondent has two sessions in the same day, which have a mix of overlapping items (e.g., things asked twice to get a response) and different items, the items will be repeated within the respondent-day dyad based on the two session starts. Figure 13 shows the distribution of relative durations after this cleaning.

Figure 13. Distribution of Item-Specific Relative Durations (Clean)



While Figure 13 shows the distribution of durations across items and responders/nonresponders, our strategy for estimating the impact or item order on whether the item had a response relies on *within-item* variation in when the item is posed to a particular respondent. More specifically, we estimate the following model with linear regression, indexing respondents with i and items with k:

$$\text{Donotrespondtoitem}(1 = \text{yes})_{ik} = \alpha + \beta_1 \text{Relativeduration}_{ik} + \gamma_i + \delta_k + \epsilon_{ik}.$$

Thus, in understanding how the time at which item k is presented to responent i affects the probability of not responding to that item, we use the respondent-specific fixed effect γ_i to hold constant the average rate at which people respond to any given item, and the item-specific fixed effect, δ_k, to hold constant the rate at which all respondents across the sample generally respond to that item.

The model, focusing on responders, thus exploits between-responder variation in when an item occurs relative to the start of the survey for different responders (e.g., due to different skip logic or whether the respondent completes the survey in one session or multiple sessions). In addition to the relative duration item, we construct the analytic sample of responder-item pairs as follows:

Restrict to responders: even though we have trace file data on both responders and some

[42] The action of "Enter Field" on STARTCP.

nonresponders, the distribution shows that, as expected, nonresponders lack a meaningful number of items with durations. In addition, since our outcome variable depends on the post-edit IUF file, we lack data on response status for those who might be classified as nonresponders due to completing very few items.

Match items between the trace file and the post-edit IUF file: since survey items differ from instrument items, we use the AHS data dictionary as a crosswalk between variable name and instrument variable name.

Code two versions of whether a person responds to a focal item: one version just contains "not reported"; another version counts nonresponse if either "not reported" or imputed on the edit flag variable. We also create a separate binary indicator for whether a responder is marked as not applicable to that item.

We then merge the information on the respondent's survey item response status to the relative duration of that item for that particular respondent[43]

Since we do not observe all items in the trace file for each responder, we create two versions of the duration variable: one with the values from above; the other that imputes respondents missing duration for a particular item to the mean duration for that item. We filter out item-respondent pairs for which the response was "not applicable," under the logic that these items might have higher missingness of durations and that not applicable might reflect skip logics rather than affirmative responses or active decisions to skip.

We estimate the regression using the `felm` function in `R's lfe` package, which helps with efficient estimation given the large number of respondent-specific fixed effects. For comparison, we also estimate models with responder fixed effects only.

If the coefficient on β_1 is significant and positive, it means that respondents are less likely to respond to items later in the survey.[44] Figure 14 shows how we have sufficient within-respondent variation in relative duration to identify an effect. Appendix Table 10 shows the 120 items included in the analysis and their mean duration. To validate a few with reference to the AHS item booklet:

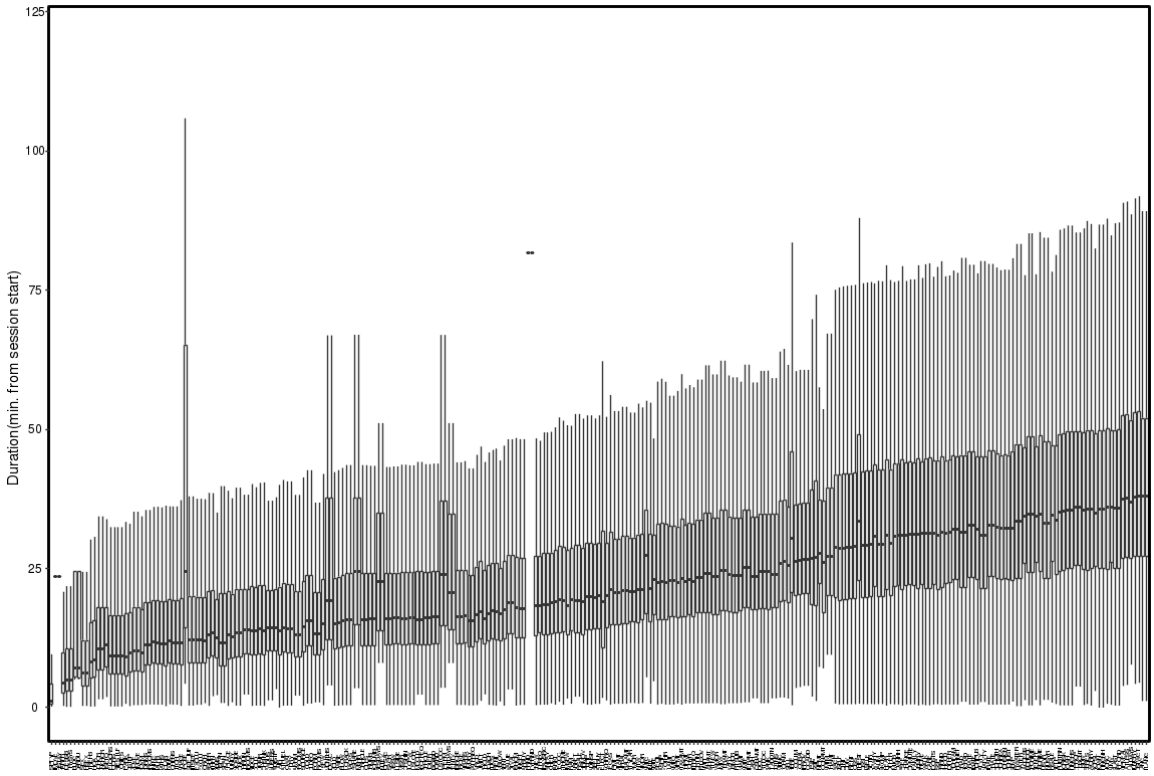
`ENTRYSYS`: whether multifamily household has entry system; ranked 2 in inferred item order and is on page 14 of the item booklet. This makes sense given items in the item booklet that are labeled in the trace file in ways that make matching with the survey instrument difficult.

`GARAGE`: presence of garage; ranked 18 in inferred item order and is on page 51 of the item booklet

`NHQSCRIME`: measure of perceptions of serious crime discussed earlier; ranked 100 in inferred item order and is on page 224 of the item booklet

[43] Since a given survey item might be comprised of multiple instruments, where that occurs, we take the max duration across instrument items for a given survey item.
[44] Since our outcome variable is nonresponse to the item.

Figure 14. Between Respondent Variation in Relative Duration for the Same Item
The x-axis contains each of the items used in the duration analysis. They are ordered by their mean duration across respondents. The box plot shows substantial between-respondent variation in when exactly the item was posed to the respondent.



Overall, the parsing shows how some items that might be more sensitive like the perceptions of neighborhood items are also toward the end of the survey instrument—but that we have sufficient between-respondent variation in item placement to look at the causal effect of ordering.

Results

We show results from two models, each with two specifications. First is a model that only includes fixed effects for the responder—this is meant to net out responder-specific propensities of skipping certain items or ending the survey at a certain point. The estimate on duration for this model thus reflects a mixture of an item's order and its intrinsic content. Second is the model specified earlier that supplements the responder fixed effects with item fixed effects—the causal effects of the item's order in this model are identified solely off of an item's relative duration for responder i compared to other responders. For instance, if two responders each receive the item about perceptions of serious crime in the neighborhood, but one responder receives the item 35 minutes into the survey based on their speed and skip logics; another responder 39 minutes into the survey, if the 39-minute respondent is less likely to respond than the 35 minute respondent, that would be evidence of an effect of duration net of the item's content and general survey placement.

Table 5 shows the results, with all models predicting *nonresponse* so a positive coefficient indicating that higher duration is associated with a higher likelihood of nonresponse. We see that results from the model with respondent fixed effects are in the expected direction—netting out

general respondent propensities to respond, we see that items with a higher relative duration are more likely to be not responded to. But the model with both item and respondent fixed effects, which analyzes order effects only off of between-respondent variation in when a particular item was reached, does not show that pattern. Further investigation is needed, but the analysis shows generally that the dual placement of potentially sensitive items at the end of the survey might lead to nonresponse due to a mix of item content and order, since the *relative* duration of those items among the same respondents does not produce results in the expected direction.

Table 5. Effect of Item Order on Item-Level Nonresponse

All models (1) subset to only respondents in the 2019 wave, (2) exclude respondent-item dyads for which "not applicable" was the variable's value. We see that imputing duration for items missing from a respondent's trace file (either actual missing or potentially due to parsing challenges) does not substantially change the results. Instead, the main change is from contrasting the respondent fixed effects model with the respondents + item fixed effects model.

Model	Treatment of Items Missing Furation	Coefficient	p-Value
Respondent FE	Listwise	0.000334500	p < 0.001
Respondent FE	Impute to mean item duration	0.000542200	p < 0.001
Respondent + item FE	Listwise	– 0.000414800	p < 0.001
Respondent + item FE	Impute to mean item duration	– 0.000274700	p < 0.001

4.2 Predicting Panel Attrition

Background

In this analysis, we leverage the longitudinal nature of the AHS to shed light on what kinds of units drop out of the panel. Specifically, we look at which variables in the 2015 AHS best predict respondent refusal in the 2017 AHS. We focus on refusal because it has a clear behavioral dimension and is the main reason for noninterviews in the 2017 AHS (70 percent of noninterviews were due to refusal).

Unlike the other analyses presented in this memo, we are able to predict refusal here using the full set of variables measured in the AHS, since we are interested in the 2017 behavior of people in units where an interview was conducted in 2015.[45] There are hundreds of categorical and numeric variables to choose from in the 2015 AHS. We therefore rely on an automated procedure that identifies the best (linear) predictors of 2017 refusal, called a penalized lasso regression (see Section 3 above for a longer description of this procedure).

Methods

We begin by restricting the sample to occupied interviews in the 2015 national survey.[46] For categorical variables, we create one dummy variable corresponding to each level, including one that indicates whether the response was missing or inapplicable for that question. Missing responses pose a bigger problem for numeric variables. For those, we impute missing values using the average of the nonmissing responses, and include in the set of potential predictors a dummy variable for each numeric variable, indicating whether mean imputation was employed.

[45] Note that this is a simplified way to refer to people within units. Given that the AHS is a survey of housing units and not households, the people residing within a unit may change between waves. Even still, the characteristics of the people responding in one wave can be predictive of the unit responding in a separate wave.

[46] For these analyses, we use the Public Use File (PUF) combined with the public case history file. This means that the universe of variables is the PUF-only variables rather than PUF variables + the IUF-only variables.

Finally, we code a variable that indicates the proportion of all predictors that were flagged as edited in the AHS (i.e., the proportion of the so-called "J" variables that was not equal to zero for a given respondent).

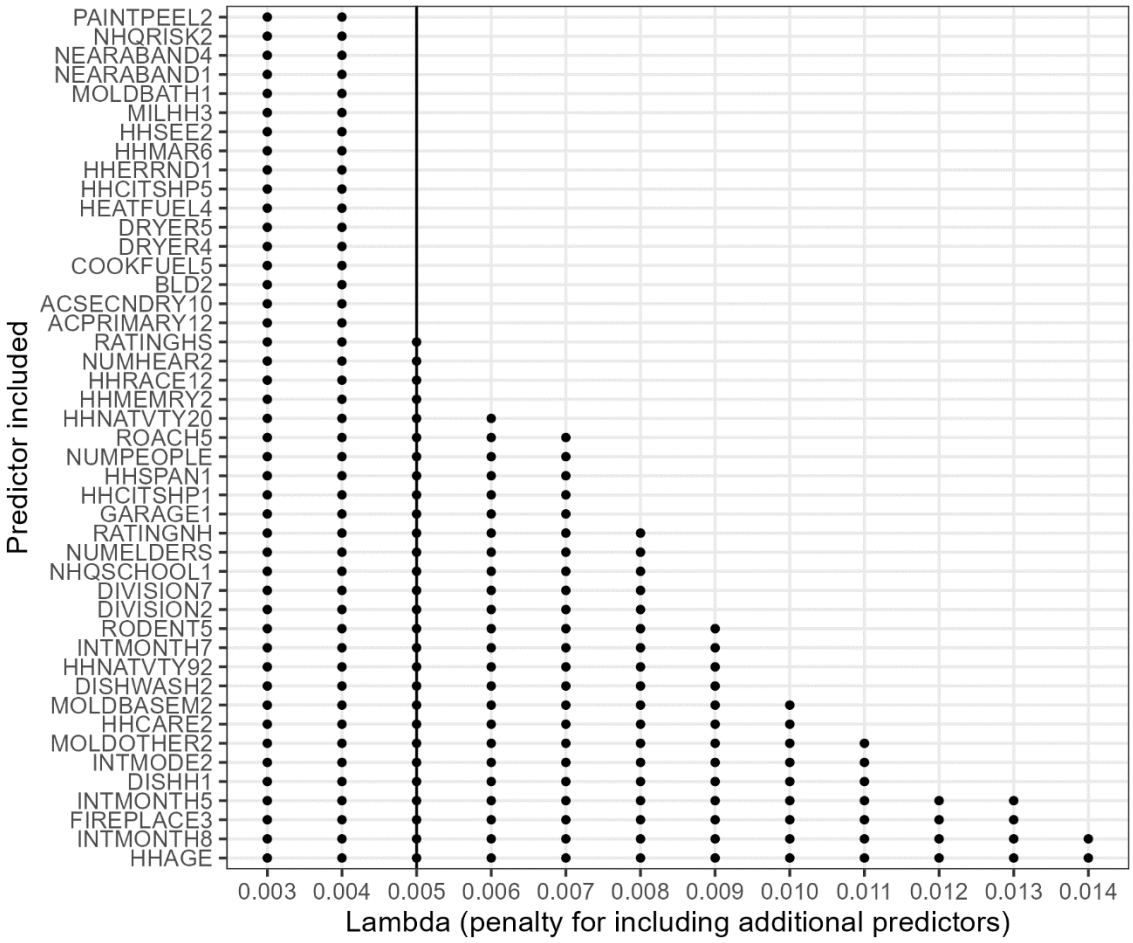
With the cleaned set of predictors in hand, we have a total of 463 possible predictors of 2017 refusal. We weight observations by the composite weight variable in all analyses, which adjusts for nonresponse bias in a given wave, but does not account for wave-on-wave patterns. We use the lasso variable selection procedure discussed in Section 3, though implement it using glmnet in R. As with all such analyses, an issue that arises is what penalty to apply to the addition of each predictor in the model. Here, we simply show how the model changes as we change the penalty (λ), and locate the penalty in a range that contains a sharp discontinuity in the number of parameters included.

Results

Figure 15 plots the different models that result from applying an increasingly stronger penalty, λ, for including additional predictors. For example, if we set $\lambda = 0.014$, the lasso regression drops all variables except INTMONTH8 (an indicator for whether the 2015 interview took place in August) and HHAGE (the age of the householder) in its search for the model that best predicts 2017 refusal. The vertical dotted line at 0.005 indicates the level of penalty chosen for this analysis, because this level appears to represent a sharp discontinuity in the number of variables selected.

Figure 15. Predictors Included in the Lasso Model as a Function of the Severity of the Penalty for Including Additional Predictors

The vertical axis lists candidate predictors of 2017 attrition contained in the 2015 AHS, in descending order of their probability of inclusion for a given penalty. The horizontal axis lists various levels of λ used to fit successive lasso regression models. As the penalty increases, so does the likelihood that variables will be excluded or "zeroed out" from the model. Each point indicates that a predictor was included at a given level of λ. The vertical line indicates the level of λ used to fit the regression model discussed further below.



Using the variables indicated as selected with points on the vertical line on Figure 15, we fit a weighted linear model predicting 2017 refusal. To estimate variance, we employ the standard replicate weights.

The first test described in our pre-analysis plan is an F-test of whether adding these variables produces a statistically significant improvement in our ability to predict 2017 refusal. Intuitively, the F-test answers the question: given that adding any variables to a model will improve its predictive accuracy just due to chance correlations, what is the probability that we would see an improvement in predictive accuracy as large as the one we do observe if none of the variables were actually related to 2017 refusal? The p-value indicates that this probability is very low ($p < .0001$). In other words, adding these 28 predictors produces a statistically significant improvement in our ability to predict refusal. This constitutes prima facie evidence that 2017 refusal is systematically related to characteristics of units measured in 2015.

When characterizing which variables do a good job of predicting whether 2015 responders drop out in 2017, some caveats are in order. First, we cannot be sure that the respondent who answered the survey in 2015 is the same person who refused the survey pertaining to that unit in 2017—it is possible that in many cases the respondent has changed, and we are predicting turnover between different residents of the same unit as much as we are predicting dropout of the same residents. Second, we must be careful in drawing causal inferences. Correlations between responses in 2015 and refusal in 2017 in many cases will arise due to unmeasured common causes.

Table 6 presents the 15 predictors that the lasso is least likely to drop as the penalty increases (e.g., those at the bottom of Figure 15).

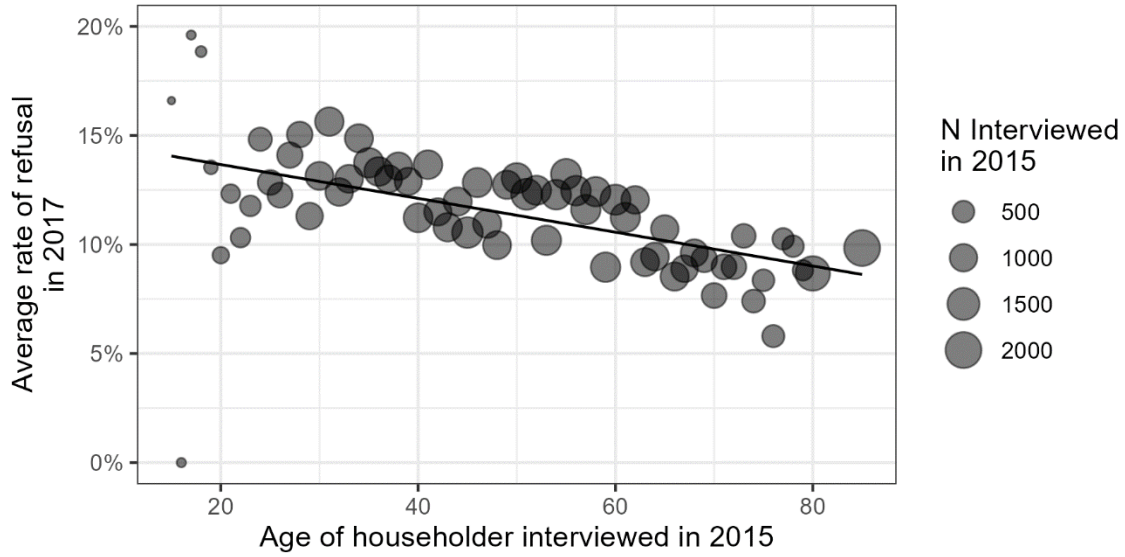
Table 6. Fifteen Predictors of 2017 Survey Refusal Among 2015 Respondents That Are Least Likely to Be Dropped by the Lasso

A subset of coefficients estimated through lasso regression. The model uses variables from the 2015 AHS to predict refusal in the 2017 wave. For a given penalty level, lasso regression selects the subset of predictors that trade off improvements in predictive accuracy against a penalty incurred by increasing the number of predictors in the model. In theory, if the penalty is set correctly, the algorithm will include the minimal subset of variables that do a good job of predicting the outcome, and will exclude those that do not add to the predictive accuracy, either because they are redundant (collinear with already included variables), highly correlated with a variable that is chosen, or do a poor job of predicting. This model uses the model corresponding to the λ penalty indicated with a vertical line on Figure 15 (0.005). Standard errors and p-values are derived from the composite replicate weights produced by the U.S. Census Bureau and employ Fay's BRR method.

Term	Estimate	Standard Error	Statistic	p-Value
HHAGE	– 0.001	0.000	– 5.926	0.000
INTMONTH8	0.040	0.006	7.076	0.000
FIREPLACE3	0.153	0.044	3.516	0.001
INTMONTH5	– 0.010	0.004	– 2.450	0.016
DISHH1	– 0.009	0.005	– 1.896	0.060
INTMODE2	– 0.011	0.004	– 2.738	0.007
MOLDOTHER2	– 0.028	0.021	– 1.319	0.190
HHCARE2	– 0.021	0.007	– 3.149	0.002
MOLDBASEM2	– 0.015	0.020	– 0.779	0.437
DISHWASH2	– 0.014	0.003	– 4.026	0.000
HHNATVTY92	– 0.010	0.008	– 1.369	0.173
INTMONTH7	0.019	0.005	4.146	0.000
RODENT5	0.025	0.005	5.235	0.000
DIVISION2	0.029	0.006	5.161	0.000
DIVISION7	– 0.019	0.005	– 3.908	0.000

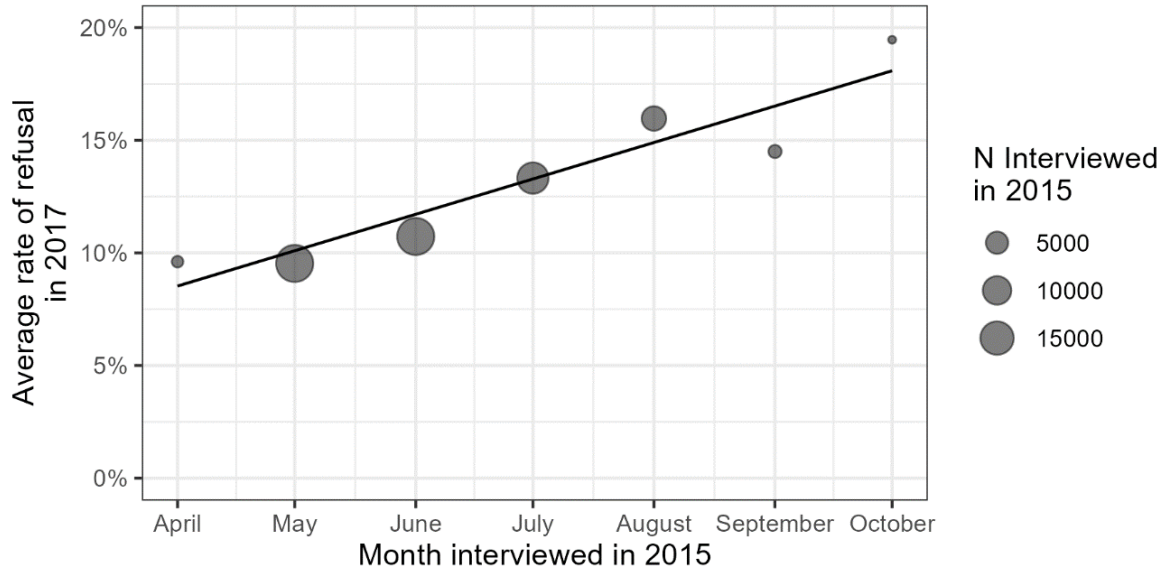
Turning to the first coefficient, `HHAGE`, we see that age is negatively correlated with the probability of refusal. As depicted on 16, the relationship is approximately linear: as the age of the householder interviewed in 2015 decreases, so too does the probability of that household refusing to do the survey in 2017.

Figure 16. Households with young respondents in 2015 are much more likely to refuse in 2017.
Each point indicates a weighted estimate of the proportion of 2017 refusers (vertical axis) for each year of age bin (horizontal axis) for householders in the 2015 AHS. The size of each point corresponds to the sample size of responders in 2015. The line is a linear least squares regression slope.



As described above, variables labeled `INTMONTH` on Table 6 are binary indicators for whether the 2015 survey was conducted in the month corresponding to the final integer. The bivariate linear relationship between 2015 interview month and 2017 refusal is depicted on Figure 17.

Figure 17. Units that were interviewed later in the 2015 round of surveying are much more likely to refuse in 2017.
Each point indicates a weighted estimate of the proportion of 2017 refusers (vertical axis) for each 2015 month of interview bin (horizontal axis). The size of each point corresponds to the sample size of responders in 2015. The line is a linear least squares regression slope.



By June, two-thirds of the 2015 sample had already been interviewed. Roughly 10 percent of those units would have refusing respondents two years later. The rate of refusal is higher for those interviewed in July, at 13 percent, but not substantially above average. Those remaining

two percent of units whose respondents were interviewed in the final months of the 2015 survey, however, exhibit a very high likelihood of 2017 refusal. One obvious explanation is that the respondents who are interviewed late in the survey are those who are the most unavailable: it is then quite unsurprising that, when those same people are sought out two years later, they are still hard to contact or just refuse to be interviewed.[47]

The coefficients on DIVISION2-7 on Table 6 indicate that 2017 refusal rates also vary by the geographic area in which the AHS is conducted. Looking at the raw data, refusal rates are highest in the Mid-Atlantic (13 percent), New England (12 percent), and East North Central (12 percent) Census divisions, and lowest in the West South Central division (9 percent).

The other coefficients do not present relationships that are quite as clear, and some appear to be the result of sparse categories happening to capture many 2017 refusers or nonrefusers.[48] Briefly, though, the lasso suggests 2017 refusal is more likely in houses that, in 2015, had: no people with disabilities living in them (DISHH1 negative); mold (MOLDOTHER2 and MOLDBASEM2 negative); householders who have difficulty dressing themselves (HHCARE2 negative); no dishwasher (DISHWASH2 negative); no problems with rodents (RODENT5 positive).

More investigation into the causes of panel attrition is encouraged. However, from this analysis, it is clear that even without a clear understanding of mechanisms, there are systematic patterns to units dropping out of the panel, which implies nonresponse bias.

4.3 Section Summary

This section explores how nonresponse bias can find its way into a sample of already responding units. Questions that are particularly sensitive—such as those pertaining to the amount of crime in the neighborhood—are most likely to go unanswered by responders. We do not find strong evidence that the placement of questions later in the survey leads to lower likelihood of answers. Turning our attention to the question of which 2015 responders drop out in 2017, we find units with younger householders interviewed later in the 2015 survey were most likely to drop out in 2017. A host of other characteristics measured in the 2015 survey are also associated with the probability of dropping out, but no clear pattern emerges.

[47] However, there are other explanations for this trend that may be worth exploring. One explanation could be a shared scheduling structure between waves—if interviewers, for instance, schedule interviews for the "inner core" of a metropolitan area first and schedule interviews for the "outer suburbs" later, it might be that units are both interviewed later in the first wave and then refuse in the later wave because they are scheduled for a time when there is less time for follow up before the end of the closeout period. Figure 18 in the appendix investigates this hypothesis, focusing on whether there is *between-region* variation in interview timing that might point to this form of confounding. The figure shows no clear differences in the distribution of 2015 interviews across months by region, which goes against the idea that respondents in certain regions are both more likely to refuse and are scheduled later. Future analyses could investigate within-region variation in scheduling as an explanation. Alternatively, refusal rates may be driven by some kind of interviewer selection, whereby interviewers put 'harder' cases lower on their list of places to visit so respondents in these units are perhaps not harder to find but were less likely to be targeted. We do not have a level of effort measure in these data and so leave this as a topic for future exploration.

[48] For example, the coefficient on FIREPLACE3 indicates that the approximately 1 percent of households whose useable fireplaces may or may not be heating equipment in 2015 are 15 percentage points more likely to refuse in 2017.

5 Consequences of Nonresponse

Stakeholders within and without government use the AHS to generate insights that can feed into important regulatory and investment decisions. This section discusses some consequences of the patterns of nonresponse analyzed in this report for applied researchers using the AHS to investigate substantive questions.

5.1 How Panel Attrition Affects Correlational Analysis

Background

If attributes of both the householder (e.g., age) and housing unit (e.g., mold, rodent infestations) in 2015 can help us predict whether or not a household refuses to be interviewed in the 2017 wave (see Section 4.2), what consequences does this entail for analyses?

One way to address this question is to investigate how attrition changes correlations that researchers might be interested in examining. For our working example, suppose a researcher is interested in examining the relationship between household income and housing inadequacy. They have a hypothesis that more affluent households are less likely to live in inadequate housing conditions. The researcher might be interested in using the multiwave structure of the AHS to assess this relationship, either to (1) increase their power to examine a relationship by pooling multiple waves, or (2) explore how the relationship changes over time (e.g., whether improved oversight of rental housing conditions might be associated with a flatter income-adequacy relationship).

Focusing on the second, if households with a certain combination of attributes is more likely to attrit than others—e.g., low-income households living in *adequate* housing being more likely to attrit than low-income households living in *inadequate* housing—this nonrandom attrition causes particular bias for investigating longitudinal trends.

Methods

To assess this form of bias, we use the Becketti, Gould, Lillard, and Welch (BGLW) pooling test to explore potential bias caused by attrition between the two panels. In the main text analysis, we focus on exploring variation between two groups: respondents interviewed in 2015 who respond in 2017 and respondents interviewed in 2015 who refuse an interview in 2017.[49] We examine the relationship between the household total income (`HINCP`) and whether the respondent lives in inadequate housing.[50] We also control for the respondent's region, which the previous section showed is a significant predictor of refusal rates.

Results

Appendix Table 13 shows the results. As expected, households with higher income are significantly less likely to live in inadequate housing conditions (negative and statistically significant coefficient). Appendix Figure 26 shows the unconditional relationship between housing adequacy and refusal, showing that a slightly higher proportion of refusers live in adequate housing. Yet, more important for this test are the interaction terms. We see that, in

[49] These are based on the NOINT variable in the case history file.

[50] More specifically, we used the ADEQUACY variable and constructed a binary measure of the unit not being adequate if the response was either moderately or severely inadequate. The regression presents income scaled by $10,000 for the purposes of interpreting coefficients; the predicted values present the nonscaled version.

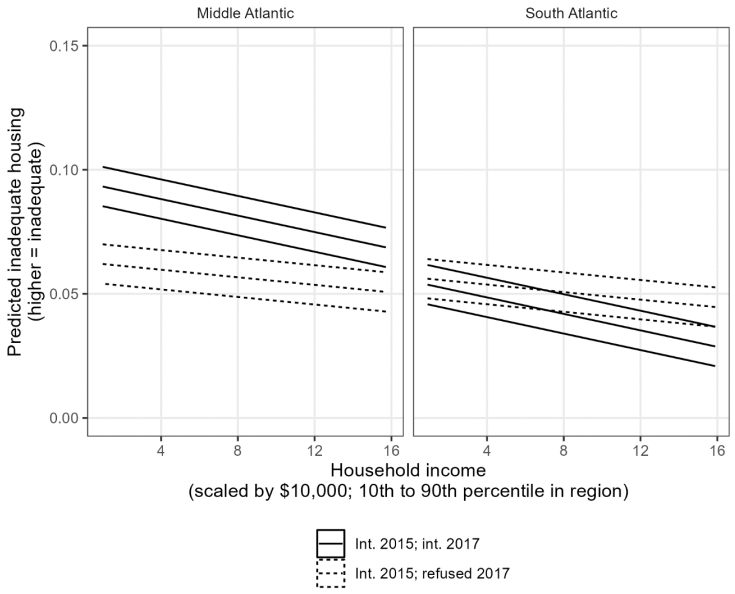
2015, respondents who go on to refuse participation in 2017 have a significantly different relationship between income and housing adequacy than those who remain in the survey (significant interaction terms). In other words, an analyst looking at this relationship using the 2017 data may get a different result if all of the units that responded in 2015 also responded in 2017.[51]

How might significant interactions between refusal and income affect the inferences an analyst makes about the relationship between income and housing adequacy? Figure 18 shows the 2015 relationship between housing adequacy (vertical axis) and income (horizontal axis) in the Middle Atlantic and South Atlantic divisions. Solid lines correspond to units who responded both in 2015 and in 2017 while dashed lines indicate those who responded in 2015 but refused in 2017. The main takeaway from this graph is that, in the Middle Atlantic division, the 2015 relationship between housing adequacy and income looks very different among those who respond in both waves compared to those who respond in 2015 only, whereas in the South Atlantic the relationship is much more similar. In the Middle Atlantic region, among those who did not attrit from the survey, there is a clear negative relationship: those with higher incomes are less likely to live in inadequate housing. Among those who attrit from the survey, the relationship is much flatter. In the South Atlantic region, there are no clear differences in adequacy between attritors and nonattritors with similar income levels. Researchers who restrict an analysis of longitudinal trends to households that appear in both waves would essentially only estimate the solid line, ignoring the dashed. This would *overstate* the relationship between income and adequacy. The composite AHS survey weights would not necessarily correct for this bias, as they do not include information on income in the reweighting scheme, and likely do not reweight for partial attrition between panels.

[51] An F-test that looks at whether adding the interaction terms between attrition and each variable produces significant improvements over a model with main effects for each variable and no interactions with attrition ($p = 0.03$).

Figure 18. Predicted Inadequacy by Income in 2015—Respondents Who Then Refuse in 2017 Versus Respondents Who Respond Both Waves

The analysis constructs a binary measure of inadequacy from the broader three-level adequacy variable. It is restricted to occupied interviews and refusers.



5.2 How Nonresponse Affects Metropolitan-Level Estimates

Background

Finally, one of the core uses of the AHS is to derive accurate metropolitan-level estimates of certain important housing stock features. Here, we investigate the extent to which 2015 AHS estimates diverge from the 2010 Decennial population count at the metropolitan area level

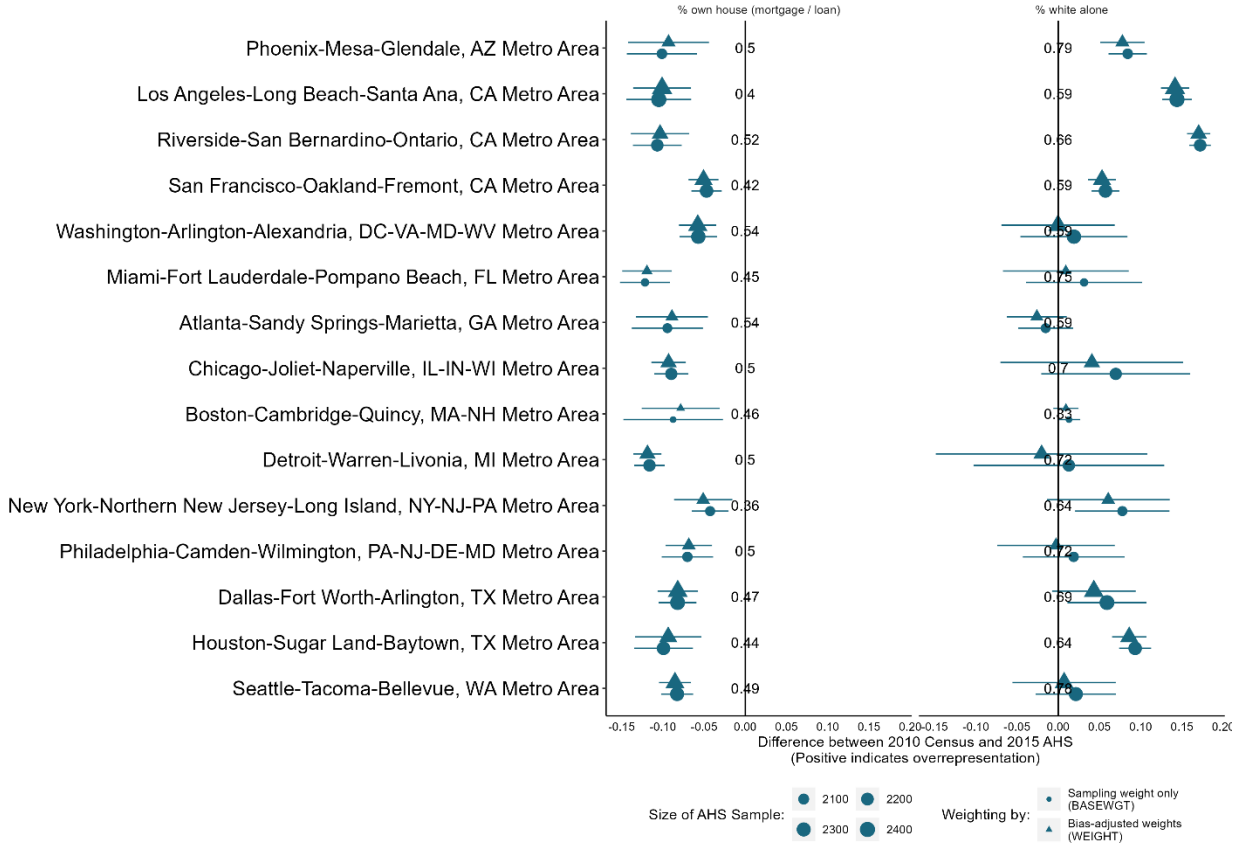
Methods

See section 2.1 above for the methods employed in the national-level benchmarking analysis. Here, we apply the same method at the metropolitan area level. For illustrative purposes, we restrict attention to two variables that appeared to diverge strongly in the national-level analysis: the proportion of householders estimated to own their house while owing a loan or mortgage and the proportion of householders who identify as white alone.

Results

The comparison of metropolitan-level divergences reveals an interesting pattern. Estimates of the proportion of householders who own their house while owing a loan or mortgage consistently underrepresent the Census count across metropolitan areas. When it comes to the count of white householders, however, the divergences vary by state. In Arizona, California, and Texas, white people are overrepresented in the AHS relative to the Decennial Census—in some cases by up to 15 percentage points—whereas in most other areas there is no statistically significant divergence. As with the prior analyses, we caution that we may be misstating the true magnitude of bias due to differential demographic changes across regions.

Figure 19. Metropolitan-Level Divergences Between 2015 AHS Estimates and 2010 Decennial Census Counts of the Proportion of Householders Who Own Their House With a Mortgage or Loan Owing and of the Proportion of Householders Who Identify as White Alone
See note on Figure 1.



5.3 Section Summary

The results of this section suggest nonresponse bias present in the AHS may affect key statistics, even with the use of weights designed to address nonresponse bias. We conducted an analysis of how panel attrition affects estimates of important correlations such as that between income and housing adequacy. Among units that responded in 2015, those that also responded in 2017 exhibit a very different relationship between income and adequacy than those who dropped out. This distinction is particularly sharp in the Middle Atlantic. Any analysis of longitudinal trends that restricted attention to units who respond in all waves of the panel would consequently overestimate the negative relationship between income and adequacy, even when employing weights. Similarly, metropolitan-level estimates from the 2015 AHS differ from the 2010 Census in ways that matter more for some regions and for some variables than for others. Whereas those who own a house with a mortgage or loan owing are consistently undercounted in all metropolitan areas, the proportion of non-White respondents is most severely undercounted in metropolitan areas located in the states of California, Arizona, and Texas.

Conclusion

This memorandum has described several methods for characterizing nonresponse bias. Among the conclusions are that: the AHS fails to reproduce population features from the 2010 Census and that the characteristics of responding and nonresponding units are different to an extent that

cannot be explained by chance. Taken as a whole, the analyses documented in this memorandum demonstrate strong evidence that nonresponse is systematically related to the characteristics of housing units and the respondents living within, which is evidence of nonresponse bias. Our analysis suggests that the nonresponse adjustment factors utilized to produce population estimates help to correct for issues of nonresponse bias, but do not completely mitigate the problem. The evidence produced in this document suggest the AHS could be strengthened with efforts designed to increase the representativeness of the responding units. This does not call for an increase in overall response rates but instead calls for efforts to increase the response rate *especially among units that are currently underrepresented*. It is encouraging that our models for predicting nonresponse perform well. This suggests that interventions can be designed to target specific units, induce a higher response rate among such units, and ultimately create a stronger, more reliable survey product.

References

Lewis, Taylor. 2015. "Replica on Techniques for Variance Approximation." SAS Support Paper, nos. 2601-2015.

Maitland, Aaron, Amy Lin, David Cantor, Mike Jones, Richard P Moser, Bradford W Hesse, Terisa Davis, and Kelly D Blake. 2017. "A Nonresponse Bias Analysis of the Health Informa on National Trends Survey (HINTS)," *Journal of Health Communication* 22 (7): 545–553.

Schouten, Barry, Fannie Cobben, Jelke Bethlehem, et al. 2009. "Indicators for the Representativeness of Survey Response," *Survey Methodology* 35 (1): 101–113.

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

A.1: Appendix

A.1 Additional Results From the Chi-Squared Analysis

Table 7. P-Values From Chi-Squared Analysis of Differences Between Responders and Nonresponders

var_tomerge	level_lab_cleaner	diffRNR_2019	p_forprint
DIVISION	East North Central	0.0060	p < 0.001
DIVISION	East South Central	0.0060	p < 0.001
DIVISION	Middle Atlantic	– 0.0169	p < 0.001
DIVISION	Mountain	– 0.0224	p < 0.001
DIVISION	New England	– 0.0147	p < 0.001
DIVISION	Pacific	0.0262	p < 0.001
DIVISION	South Atlantic	0.0404	p < 0.001
DIVISION	West North Central	– 0.0086	p < 0.001
DIVISION	West South Central	– 0.0161	p < 0.001
FL_SUBSIZ	No	– 0.0016	p < 0.001
FL_SUBSIZ	Yes	0.0016	p < 0.001
HUDSAMP	No	– 0.0016	p < 0.001
HUDSAMP	Yes	0.0016	p < 0.001
METRO_2013	Metro, Central City	– 0.0033	p < 0.001
METRO_2013	Metro, Not Central City	– 0.0028	p < 0.001
METRO_2013	Micropol.	– 0.0074	p < 0.001
METRO_2013	Non Micropol.	0.0135	p < 0.001
REGION	Midwest	– 0.0026	p < 0.001
REGION	Northeast	– 0.0316	p < 0.001
REGION	South	0.0304	p < 0.001
REGION	West	0.0039	p < 0.001
RENTSUB	Missing	0.0087	p < 0.001
RENTSUB	No rental subsidy or reduction	– 0.0967	p < 0.001
RENTSUB	Other government subsidy	0.0174	p < 0.001
RENTSUB	Public housing	0.0247	p < 0.001
RENTSUB	Rent reduction	0.0056	p < 0.001
RUCC_2013	Completely rural; metro adj.	0.0013	p < 0.001
RUCC_2013	Completely rural; nonmetro adj.	– 0.0009	p < 0.001

var_tomerge	level_lab_cleaner	diffRNR_2019	p_forprint
RUCC_2013	Metro. county (<250k)	– 0.0046	p < 0.001
RUCC_2013	Metro. county (1+ mil)	– 0.0115	p < 0.001
RUCC_2013	Metro. county (250k–1mil)	0.0101	p < 0.001
RUCC_2013	Urban <20k; metro adj.	0.0124	p < 0.001
RUCC_2013	Urban <20k; nonmetro adj.	– 0.0050	p < 0.001
RUCC_2013	Urban 20k+; metro adj.	– 0.0001	p < 0.001
RUCC_2013	Urban 20k+; nonmetro adj.	– 0.0016	p < 0.001
SPSUTYPE	Not self-rep	0.0136	p < 0.001
SPSUTYPE	Self-rep	– 0.0136	p < 0.001
WPSUSTRAT	CI record	0.0000	p < 0.001
WPSUSTRAT	HUD records	0.0026	p < 0.001
WPSUSTRAT	Mobile home	– 0.0011	p < 0.001
WPSUSTRAT	Other	– 0.0207	p < 0.001
WPSUSTRAT	Other	– 0.0050	p < 0.001
WPSUSTRAT	Owners; 1 unit	0.0172	p < 0.001
WPSUSTRAT	Owners; 2+ unit	0.0037	p < 0.001
WPSUSTRAT	Renters; 1 unit	– 0.0013	p < 0.001
WPSUSTRAT	Renters; 2+ unit	0.0000	p < 0.001
WPSUSTRAT	Vacant; 1 unit	0.0037	p < 0.001
WPSUSTRAT	Vacant; 2+ unit	0.0009	p < 0.001

Note: All differences are significant at the $\mathbf{p < 0.001}$ level.

Figure 20. Differences Between Responders and Nonresponders: 2017 Wave

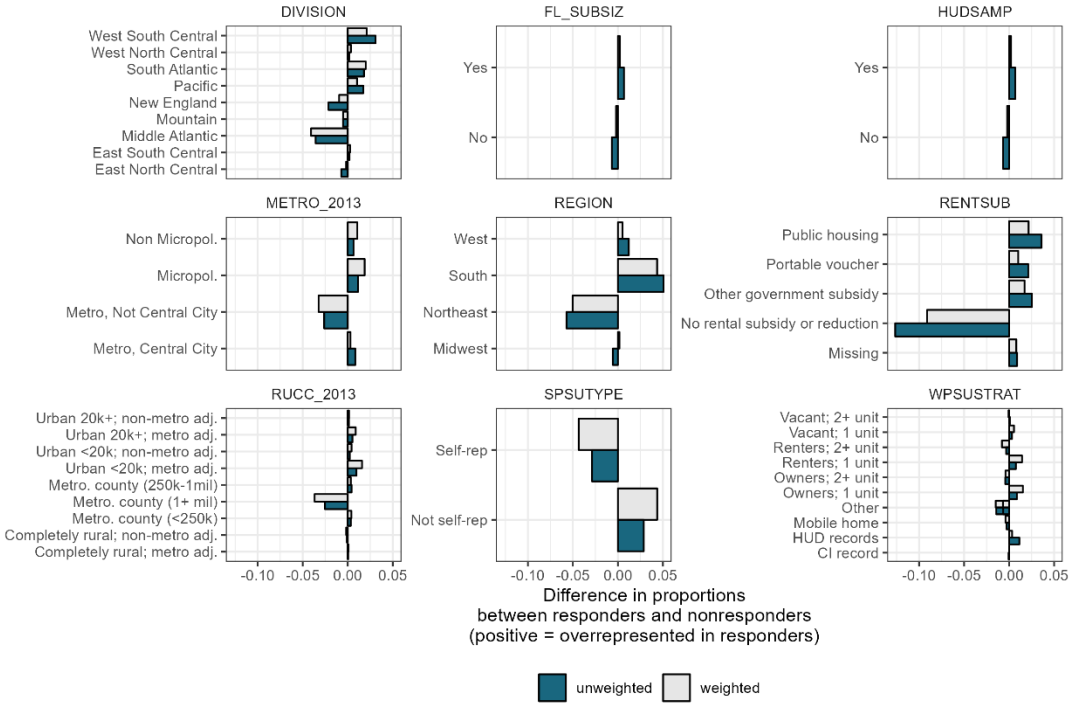
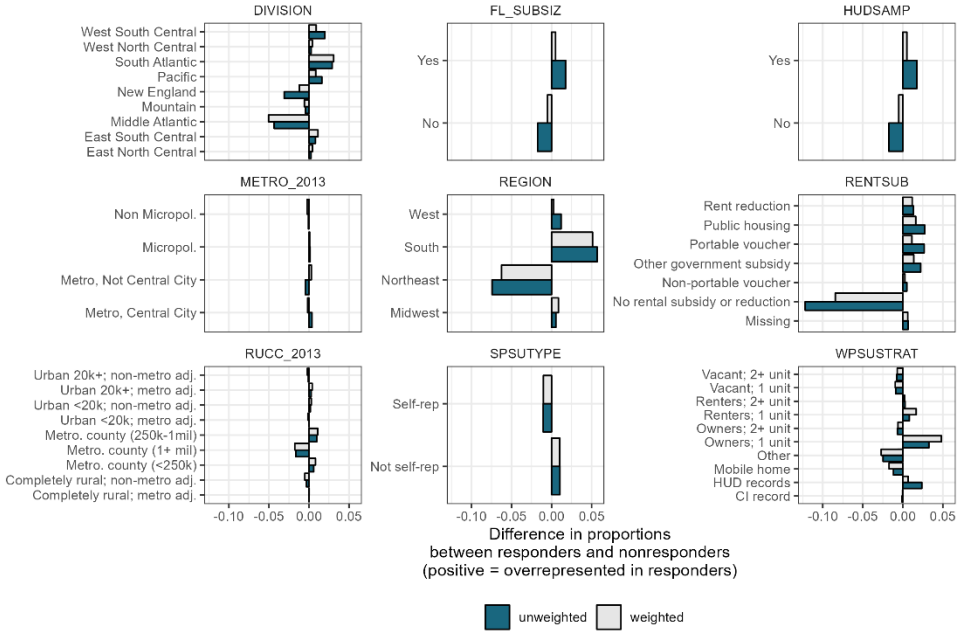


Figure 21. Differences Between Responders and Nonresponders: 2015 Wave



A.2 Additional Results From R-Indicator Analysis

Table 8. Results From R-Indicator Analysis

Wave	Estimated R	Permutation p-Value	LRT Statistic	LRT p-Value
2015	0.90	0	1291	0.00
2017	0.92	0	3228	0.00
2019	0.90	0	5964	0.00

A.3 Additional Results From the Predicting Nonresponse and Refusal Analysis

Table 9. Tract-Level Predictors From the American Community Survey

The first set of predictors (ACS 100 to 199, 1500 to 1999, etc.) represent monthly housing costs. Other predictors reflect race/ethnicity, educational attainment, and housing costs as a proportion of income.

Feature
acs_100_to_199_prop
acs_1500_to_1999_prop
acs_200_to_299_prop
acs_2000_or_more_prop
acs_300_to_399_prop
acs_400_to_499_prop
acs_500_to_599_prop
acs_600_to_699_prop
acs_700_to_799_prop
acs_800_to_899_prop
acs_900_to_999_prop
acs_asian_alone_prop
acs_at_or_above_150_percent_of_the_poverty_level_prop
acs_bachelors_degree_or_higher_prop
acs_black_or_african_american_alone_prop
acs_estimate_median_age_total
acs_estimate_median_household_income_in_the_past_12_months_in_2014_inflation_adjusted_dollars
acs_foreign_born_noncitizen_prop
acs_hispanic_or_latino_prop
acs_in_the_labor_force_unemployed_prop
acs_less_than_100_prop
acs_less_than_high_school_graduate_prop
acs_living_in_household_with_ssi_orsnap_prop
acs_native_hawaiian_and_other_pacific_islander_alone_prop
acs_owner_occupied_housing_units_zero_or_negative_income_prop
acs_renter_occupied_housing_units_20000_to_34999_20_to_29_percent_prop

Feature
acs_renter_occupied_housing_units_20000_to_34999_30_percent_or_more_prop
acs_renter_occupied_housing_units_35000_to_49999_20_to_29_percent_prop
acs_renter_occupied_housing_units_35000_to_49999_30_percent_or_more_prop
acs_renter_occupied_housing_units_50000_to_74999_20_to_29_percent_prop
acs_renter_occupied_housing_units_50000_to_74999_30_percent_or_more_prop
acs_renter_occupied_housing_units_75000_or_more_20_to_29_percent_prop
acs_renter_occupied_housing_units_75000_or_more_30_percent_or_more_prop
acs_renter_occupied_housing_units_less_than_20000_20_to_29_percent_prop
acs_renter_occupied_housing_units_less_than_20000_30_percent_or_more_prop
acs_renter_occupied_housing_units_zero_or_negative_income_prop
acs_some_college_or_associates_degree_prop
acs_some_other_race_alone_prop
acs_two_or_more_races_prop
acs_unweighted_sample_count_of_the_population

Figure 22. Ability to Predict Nonresponse: 2017 Wave
The figure shows F1 scores for two types of feature sets: AHS-only (which includes both sampling frame variables and lagged response/contact attempt variables) and those plus the ACS contextual features.

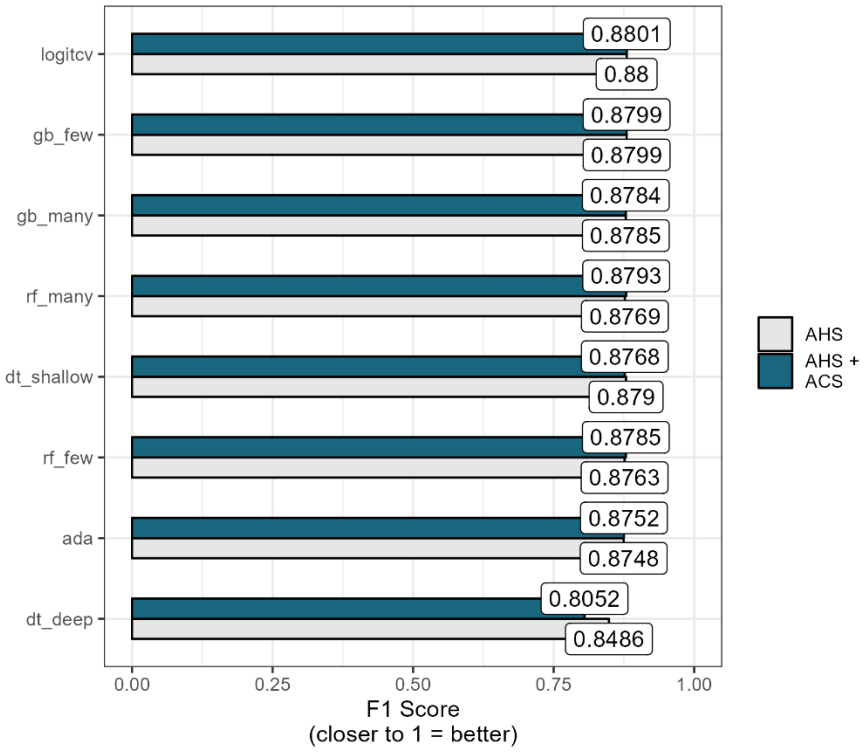


Figure 23. Ability to Predict Refusal: 2017 and 2019 Wave

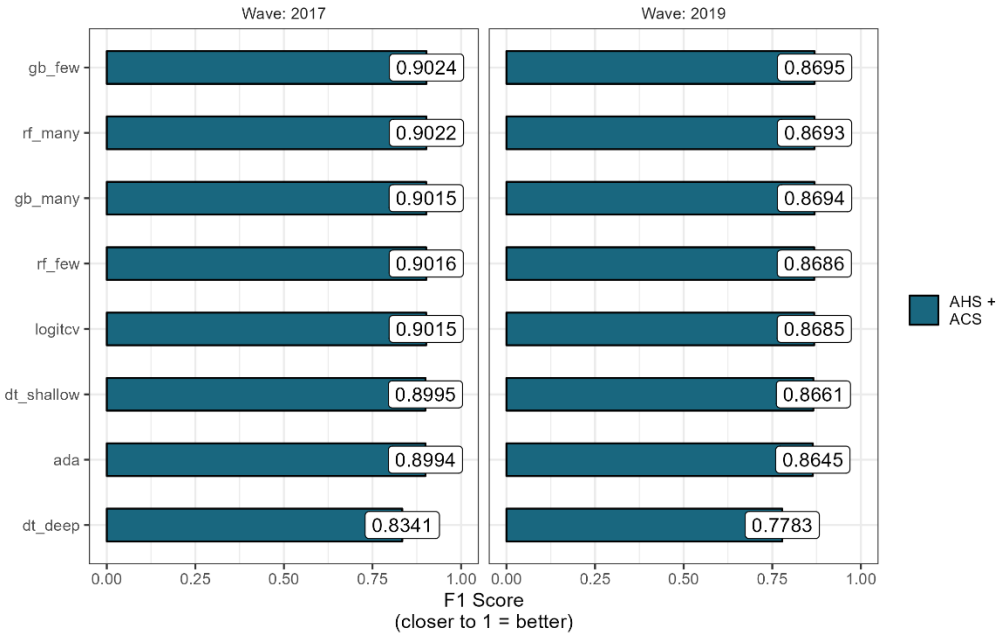


Figure 24. Remaining Feature Importance Outside of the top 20—Random Forest: 2019 Wave

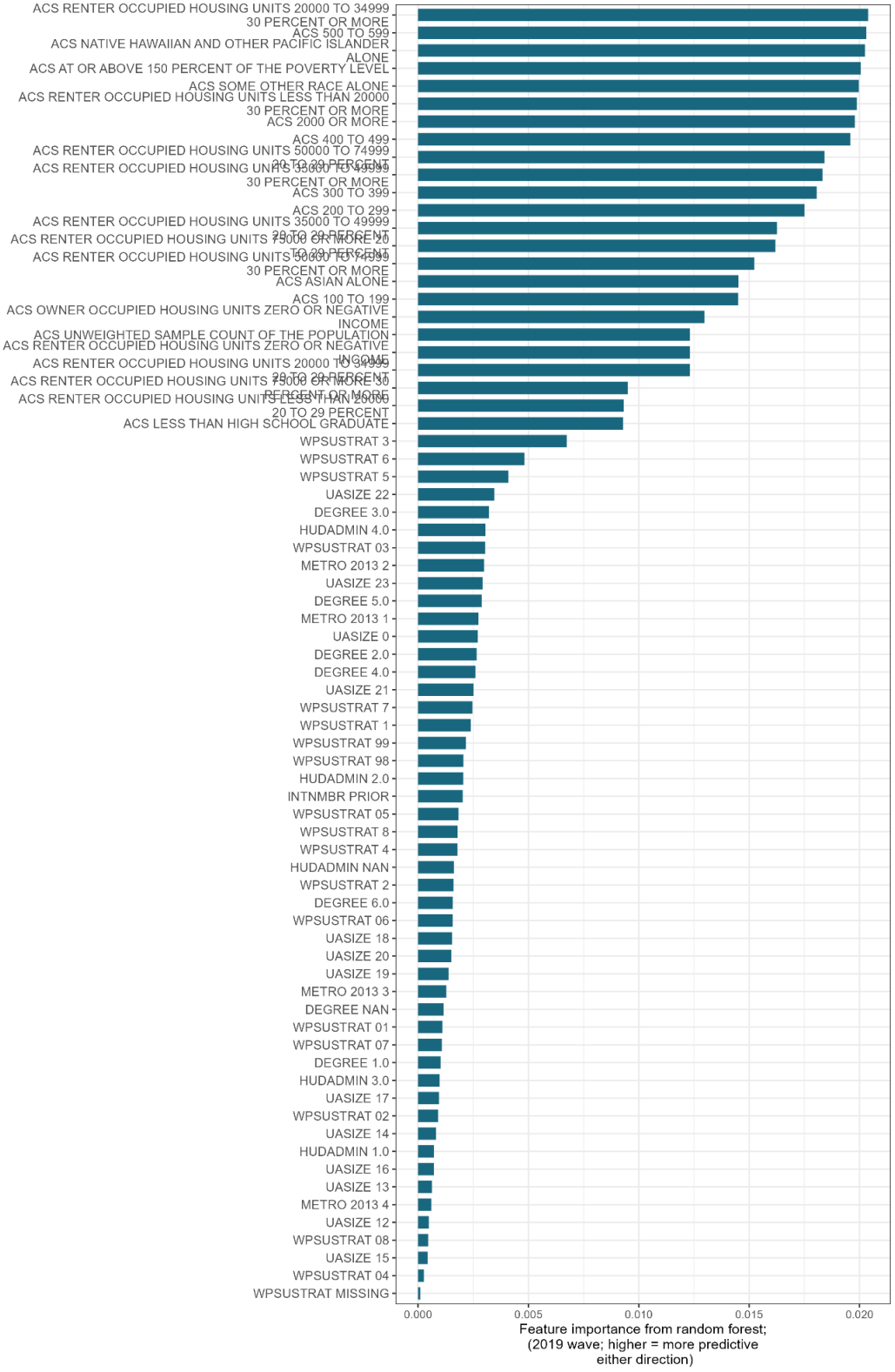


Table 10. Items in Order Effects Analysis Based on Trace File

The items are ordered by their average duration across respondents.

Variable	Average Duration	Variable	Average Duration
MHMOVE	10.17	ROOFHOLE	21.45
ENTRYSYS	11.83	WINBOARD	21.49
HHSEX	12.81	MONLSTOCC	21.50
GUTREHB	15.07	VACRNTDAYS	21.58
HOA	15.29	WINBROKE	21.60
MHANCHOR	15.40	WINBARS	21.63
STORIES	15.89	NOWATFREQ	21.66
STORIES_IUF	15.89	PLUGS	22.12
UNITFLOORS	15.89	RENT	22.39
TPARK	15.90	LOTVAL	22.49
NOSTEP	16.21	YEARBUY	22.68
HHMOVE	17.05	SUITYRRND	22.94
BEDROOMS	17.37	DWNPAYSRC	23.35
DINING	17.52	FIRSTHOME	23.96
SOLAR	17.61	FORSALE	24.25
LIVING	17.70	LEADINSP	24.34
KITCHENS	17.71	OILAMT	27.06
GARAGE	17.74	PROTAXAMT	27.73
UNITSIZE	17.91	TRASHAMT	28.28
UNITSIZE_IUF	17.91	WATERAMT	28.42
KITEXCLU	18.01	OTHERAMT	29.02
PORCH	18.03	MOVWHY	34.15
WASHER	18.45	RMJOB	34.32
OTHFN	18.48	RMOWNHH	34.48
DENS	18.51	RMCHANGE	34.61
HHSPAN	18.57	RMCOMMUTE	34.63
FIREPLACE	18.75	RMFAMILY	34.65
LAUNDY	18.82	RMHOME	34.82
MONOXIDE	19.10	RMCOSTS	34.98
UFINROOMS	19.10	RMHOOD	35.03
FRIDGE	19.15	RMOTHER	35.08
COLD	19.20	SEARCHFAM	35.25
KITCHSINK	19.28	SEARCHNET	35.38
SEWUSERS	19.34	HHGRAD	35.42
HEATFUEL	19.69	NRATE	35.63
FAMROOMS	19.74	HHNATVTY	35.68
NOWAT	19.74	SEARCHPUB	35.68
WATSOURCE	19.81	HRATE	35.85
COLDEQ	20.37	SEARCHOTH	35.86
RECROOMS	20.42	SEARCHREA	35.88
VACMONTHS	20.51	SEARCHLIST	35.93
NOWIRE	20.67	SEARCHSIGN	35.93

Variable	Average Duration	Variable	Average Duration
WALLCRACK	21.03	NEARWATER	36.53
FLOORHOLE	21.15	HMRACCESS	37.47
TIMESHARE	21.15	NHQPCRIME	37.51
FNDCRUMB	21.26	NHQSCHOOL	37.51
VACRESDAYS	21.26	HMRENEFF	37.56
ROOFSAG	21.27	NHQSCRIME	37.69
WALLSIDE	21.31	HHINUSYR	37.70
ROOFSHIN	21.34	HMRSALE	37.75
WALLSLOPE	21.37	NHQPUBTRN	37.97
COLDEQFREQ	21.41	NHQRISK	38.08
		RATINGHS	38.13
		WATFRONT	38.16
		SUBDIV	38.18
		AGERES	38.20
		FSWORRY	38.30
		CROPSL	38.40
		RATINGNH	38.43
		NORC	38.53
		FSLAST	38.73
		FSAFFORD	38.85
		FSSKIPMEAL	40.59
		FSEATLESS	40.73
		FSMEALDAYS	40.76
		FSHUNGRY	40.82
		FSLOSTWGT	40.82
		INTLANG	41.13

A.4 Item Order Effects: Additional AnalysesA.5 Predicting Panel Attrition: Additional Analyses

Table 11. Examples of Variables Removed During LASSO Preprocessing

Step	cols_removed	example_vars_removed
Edit flag variables (J variables)	312	JNOTOIL; JNUMADULTS; JHHINUSYR; JVACRNTDAYS; JRMCHANGE
High NA (over 20% missing)	196	SP1REPWGT68; HHYNGKIDS; PLUGS; RATINGNH; SP2REPWGT137

Figure 25. Rate of Refusal in 2017 by Month and Division of Interview in 2015

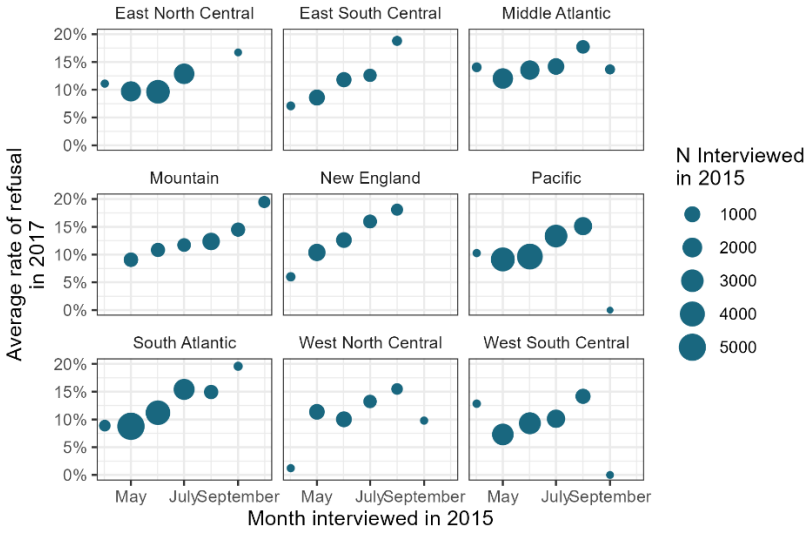


Table 12. Twenty-Five Next Strongest Predictors of 2017 Survey Refusal Among 2015 Respondents

Term	Estimate	Standard Error	Statistic	p-Value
FIREPLACE3	0.153	0.044	3.516	0.001
HHCARE2	– 0.021	0.007	– 3.149	0.002
HHCITSHP1	0.014	0.005	2.929	0.004
ROACH5	0.013	0.004	2.873	0.005
INTMODE2	– 0.011	0.004	– 2.738	0.007
NHQSCHOOL1	– 0.009	0.004	– 2.169	0.032
HHNATVTY20	0.154	0.052	2.965	0.004
HHNATVTY92	– 0.010	0.008	– 1.369	0.173
RATINGNH	– 0.002	0.001	– 2.070	0.041
NUMHEAR2	– 0.012	0.006	– 2.041	0.043
NUMELDERS	– 0.005	0.003	– 1.912	0.058
DISHH1	– 0.009	0.005	– 1.896	0.060
RATINGHS	– 0.002	0.001	– 1.834	0.069
HHMEMRY2	– 0.011	0.007	– 1.627	0.106
MOLDOTHER2	– 0.028	0.021	– 1.319	0.190

A.6 Attritor Heterogeneity: Additional Analyses

Figure 26. Adequacy Across 2017 Refusers and Nonrefusers
The proportions reweight using the composite weight.

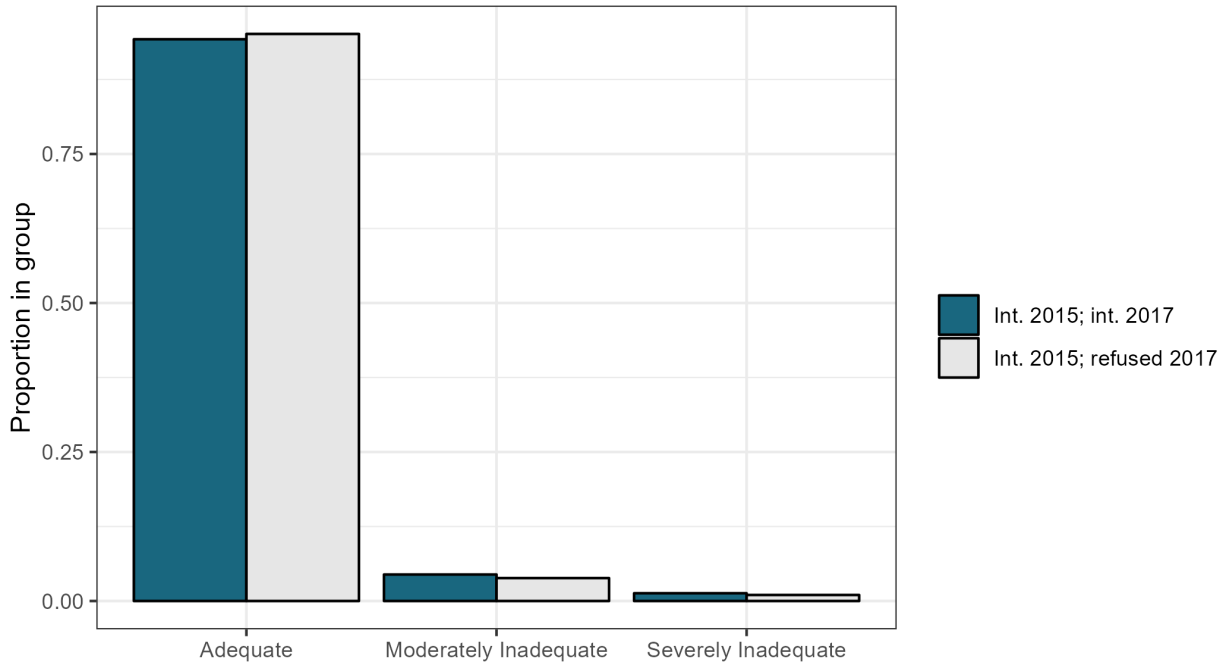


Table 13. Attritor Heterogeneity in Relationship Between Income and Adequacy—Refusers in 2017: Regression
The table shows that in addition to main relationships where those with higher incomes are less likely to have inadequate housing, we see heterogeneity in this income-adequacy relationship between attritors and nonattritors.

	Dependent Variable:
	Yes inadequate
division_descriptiveMiddle Atlantic	0.012
	(0.008)
division_descriptiveEast North Central	– 0.021 ***
	(0.007)
division_descriptiveWest North Central	– 0.018 **
	(0.009)
division_descriptiveSouth Atlantic	– 0.028 ***
	(0.007)
division_descriptiveEast South Central	– 0.003
	(0.008)
division_descriptiveWest South Central	– 0.001
	(0.008)
division_descriptiveMountain	– 0.029 ***

	Dependent Variable:
	Yes inadequate
	(0.008)
division_descriptivePacific	−0.017**
	(0.007)
refusal_17	−0.032**
	(0.014)
inc_scaled	−0.002***
	(0.0001)
division_descriptiveMiddle Atlantic:refusal_17	−0.0003
	(0.016)
division_descriptiveEast North Central:refusal_17	0.023
	(0.015)
division_descriptiveWest North Central:refusal_17	0.026
	(0.018)
division_descriptiveSouth Atlantic:refusal_17	0.033**
	(0.014)
division_descriptiveEast South Central:refusal_17	0.001
	(0.018)
division_descriptiveWest South Central:refusal_17	0.004
	(0.016)
division_descriptiveMountain:refusal_17	0.023
	(0.017)
division_descriptivePacific:refusal_17	0.014
	(0.015)
refusal_17:inc_scaled	0.001*
	(0.001)
Constant	0.083***
	(0.007)
Observations	60,487
Log Likelihood	−5,077.268
Akaike Inf. Crit.	10,194.540

Note: *p < 0.1; **p < 0.05; ***p < 0.01.

Appendix B: Design Document Project Design

Project title: Using Incentives to Reduce Nonresponse Bias in the American Housing Survey (AHS)
Project code: 1901

1 Project Objectives

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.[52]

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 overreliance on post-survey adjustment procedures. The use of nonresponse adjustment weights can add noise to population estimates, even when recovering population estimates that are accurate. By improving the quality of the data collection prior to nonresponse adjustment, we may be able to generate more precise estimates. 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 effective at increasing data quality.

In this project, we distinguish between nonresponse bias, on the one hand, and survey representativeness, on the other hand. Nonresponse bias is 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.[53] 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 nonresponse adjustment factor (NRAF) 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, 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

[52] The response rates for the 2015 and 2017 waves are taken from the AHS public methodology reports. The response rate for the 2019 wave is taken from our analysis of the IUF with the below restrictions to the national sample and excluding the bridge sample, with values based on the coding responders as STATUS == 1, 2, or 3 (n = 63,186) and nonresponders as STATUS == 4 (n = 22,965). These may differ from those in the published methodology report if there are different inclusion criteria for the published rates to remove ineligible households.
[53] In other words, it is a correlation between the propensity to respond to the survey and a key outcome of interest.

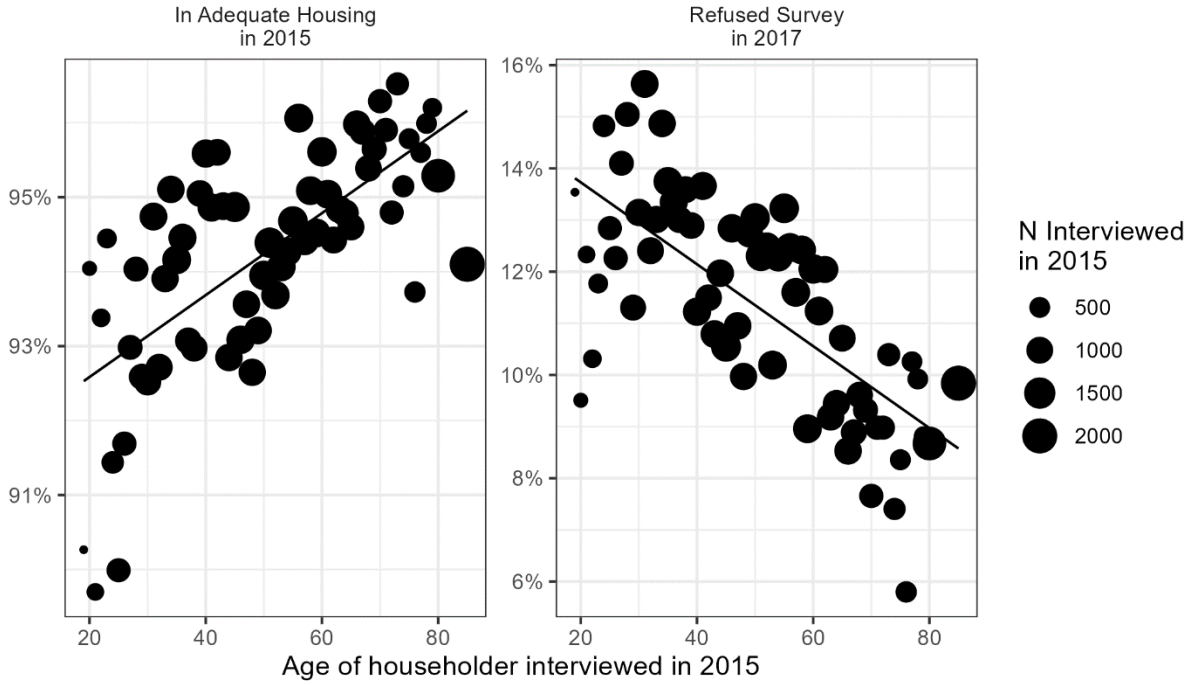
such bias adjustment schemes typically increase variance in estimates: they essentially require units in grid cells with a lot of missingness to "represent" more unobserved units than those in grid cells with less missingness.

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. For example, a key outcome the AHS measures is housing inadequacy. Among units where an interview was successfully conducted during 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. In a separate memo on nonresponse bias in prior rounds of the AHS (see attached), we found numerous systematic patterns in panel attrition whose statistical and substantive significance persists in spite of weighting meant to account for nonresponse bias. 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.[54] Attributing such divergence to nonresponse bias with complete certainty is a challenging task since, by definition, we cannot measure the outcomes of those who do not respond. However, the many pieces of evidence presented in the nonresponse bias memo suggest that, in addition to adjusting sample estimates on the backend, improving sample composition on the frontend would increase their accuracy.

The question of survey representativeness relates closely to that of nonresponse bias: it describes systematic differences between sampled units who do and do not respond to the survey on demographic and administrative variables, rather than on key outcomes. While demographic and administrative measures may often be of secondary importance to decisionmaking, they help to understand the extent to which missingness due to nonresponse is random or systematic. In our separate memo, we find responders and nonresponders differ systematically on a range of attributes, both within and between waves of the survey. These divergences are important to understand for at least three reasons: (1) demographic and administrative variables often define subgroups among whom key outcomes are estimated (e.g., the rate of housing inadequacy in rural versus urban areas); (2) as described above, these variables are employed to conduct reweighting as they are often the only ones available for nonresponders; (3) demographic and administrative variables provide a window onto nonresponse bias as they are correlated with key outcomes. See on this last point, for example, Figure 1, which illustrates that panel attrition in 2017 is predicted by the age of the householder interviewed in 2015, and that householder age is also correlated strongly with measures of housing adequacy. As such, improving the representation of units with young householders may reduce bias in estimates of housing adequacy.

[54] Significant at the $\alpha = .01$ level, using replicate weights to estimate variance.

Figure 1. Units with young householders in 2015 were (a) less likely to be adequate housing in 2015 and (b) more likely to drop out of the panel due to refusal in 2017.
Points represent reweighted estimates of proportions for different ages, size corresponds to number of respondents in 2015.



The purpose of this project is to determine whether and how the provision of cash incentives prior to contact with Census Bureau staff can achieve two related goals: reducing nonresponse bias in (adjusted and unadjusted) sample estimates and increasing representativeness of the sample. The use of incentives by Federal agencies has raised a variety of concerns about their cost, the proper use of taxpayer funds, impact on other surveys, conditioning the expectations of respondents, and implications for the "social contract" between the Federal Government and citizens. With those concerns in mind, this test of incentives is intended to generate actionable evidence on the optimal way to target incentives—both how much and to whom—in a way that maximizes data quality while minimizing the allocation of incentives to units that either are not likely to be converted to a response with an incentive (or incentive of a certain amount) or would still respond in the absence of a monetary incentive.

Because the AHS is a panel survey of housing units, we are able to take advantage of a rich set of longitudinal data not available in other surveys to improve the quality of the predictive models. In particular, in addition to the sampling frame data, we are able to include response outcomes (i.e., whether or not the unit responded) and paradata (which include the number, type, and timing of contact attempts and reasons for refusing the survey) in the 2015, 2017, and 2019 AHS. We additionally leverage time-varying neighborhood characteristics from respective American Community Survey (ACS) 5-year estimates (2014; 2016; 2018) that capture aggregate demographic characteristics (age; employment status) potentially related to nonresponse. These data sources lead to a high-dimensional dataset with 400+ predictors; we use machine learning classifiers to retain this high-dimensional predictor set in the predictions of nonresponse.

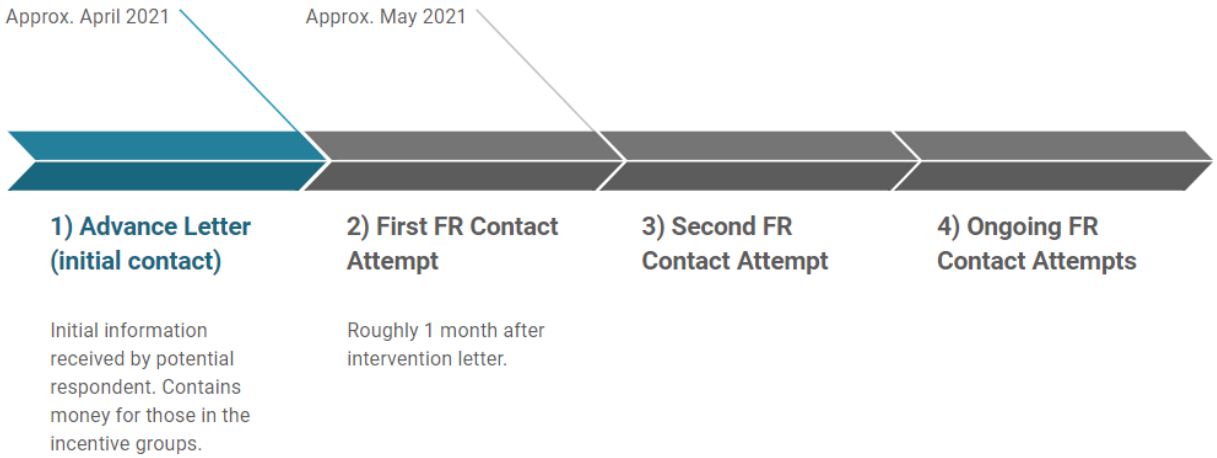
While providing large incentives to all housing units in the sample could conceivably increase both the response rate and data quality, the goal of the project is not to test the effectiveness of blanket incentives. Rather, the study is designed to generate evidence about the effectiveness of targeting incentives to different types of units with the aim of efficiently using incentives to convert the subset of important cases that would not participate in the survey absent an incentive. The goal is to move away from a uniform allocation of incentives, which is inefficient in providing incentives both to cases which are unlikely to be affected by incentives and to cases which are unlikely to introduce bias.

We expect the results to be most informative for the use of targeted incentives in future iterations of the AHS. Not all surveys are able to take advantage of the rich data available to the AHS, and lessons learned from the AHS may not be applicable to surveys with different substantive focus and/or different target populations.

2 Intervention Design

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 is delivered inside an envelope containing a letter reminding the potential respondent about the survey. This letter is 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. The timeline is depicted on Figure 2.

Figure 2. Intervention Timeline



Given the risks survey nonresponse raises—sample size reduction and possible bias—it is not surprising that a large literature has developed seeking to understand and reduce nonresponse. This project builds on a branch of this literature demonstrating the effectiveness of cash incentives at increasing response rates. We focus here on "noncontigent" and "nondiscretionary" cash incentives (Jackson, McPhee, and Lavrakas, 2020). The cash incentives are noncontingent because they are provided to respondents in advance of the survey rather than only provided upon survey completion.[55] Second, the presence and magnitude of the incentive is

[55] These are often described as "unconditional" incentives.

nondiscretionary because it is determined centrally for all survey respondents, rather than at the discretion of individual field staff for particular respondents.

In the context of the AHS, three questions are of central interest:

1. What contributes to survey nonresponse?
2. Given those contributors, to whom should surveyors allocate incentives in order to reduce nonresponse bias?
3. What magnitude of incentives should surveyors allocate?

We provide a brief overview of existing research in each area, and discuss gaps the present experiment aims to fill.

2.1 What contributes to survey nonresponse?

Groves, Singer, and Corning (2000) suggests that a lack of awareness or salience may contribute to nonresponse, while Hidi and Renninger (2006) and Ariely, Bracha, and Meier (2009) focus on lack of interest and motivation as behavioral explanations for nonresponse. In the context of a survey fielded by the federal government, distrust of government may also play a role. Certain groups may also have schedules and behavioral patterns that make them harder to contact than other groups. Our analyses suggest, for example, that units in the AHS with younger householders interviewed in 2015 were more likely to refuse in 2017.

In addition to household characteristics, the mode of surveying also appears to matter. Laurie and Lynn (2008) note that incentives are more effective in non-in-person surveys (2009: 207), possibly because of the already high response rates of in-person surveys. In the context of the AHS, the rate of telephonic surveying has increased substantially: from 27 percent in 2015, 30 percent in 2017, to 37 percent in 2019. This trend may thus have provided conditions that are particularly suited to the use of incentives, though it should be noted that the evidence on how survey mode influences incentive effectiveness is mixed.

2.2 To whom should surveyors provide incentives?

A large body of research has found that incentives *generally work* to improve response rates, regardless of a particular household's constraints and barriers to survey participation. In a meta-analysis of 49 studies, noncontingent financial incentives were predicted to increase response rates from an average of a rate of 85 percent to an average of 92 percent (Edwards et al., 2002). In a meta-analysis of over 20 years of articles, Mercer et al. (2015: 122) find that the largest marginal gains occur between $0 and $1, and taper off considerably after $2.

Yet, the bulk of the studies in these meta-analyses use the following procedure:

- Decide on an incentive amount to vary (e.g., $1 versus $5, with Mercer et al. (2015)'s review of studies showing incentives that vary between $0 and $50).
- Randomly assign sampled units to receive different incentive amounts.

While this procedure allows researchers to assess the impact of different incentive magnitudes, it ignores the fact that households differ in three ways. First is the household's likelihood of nonresponse. Second, among the pool of households with a low likelihood of response, is the extent to which that household's nonresponse contributes to bias. Third, among the pool of households with both a low likelihood of response and a high potential for that nonresponse to

contribute to bias, is the extent to which that household is likely to be impacted by incentives. A growing set of literature seeks to (1) identify these three groups, and (2) test approaches that target incentives on the basis of group membership.

Researchers affiliated with the National Center for Educational Statistics (NCES) have explored these approaches with various surveys. Crissey, Christopher, and Socha (2015), focusing on the 2013 update to the High School Longitudinal Study (HSLS) and the 2014 followup to the Beginning Postsecondary Students Longitudinal Study 2012 (BPS), estimate what they call "importance scores." The importance scores are a function of two components. First is a propensity model for nonresponse, estimated using paradata prior to the survey collection. Second is what the authors call a "bias-likelihood score," or the extent to which that nonresponse will contribute to bias. The authors estimate this score *during* data collection by finding the Mahalanobis distance along various attributes between (1) nonrespondents and (2) those that have responded thus far. The importance score is a dual function of these two inputs.

Selecting respondents with the highest importance scores, the researchers randomly allocated the magnitude of incentive promised to survey respondents if they completed the survey (contingent incentive).[56] The study introduces an important conceptual approach to targeting—first, that incentives can be targeted to a subset of respondents and second, that researchers should take into account both response propensities and contributions to bias when selecting that subset. However, by giving incentives to *all* high importance respondents, it does not causally test whether targeting represents an improvement over randomly allocated incentives—the use of targeting as such is not evaluated. Similarly, other studies investigate different ways of operationalizing whom to target with incentives—for instance, Link and Burks (2013) compare response propensities estimated using different types of variables available in address-based sampling; Coffey and Zotti (2015) combine response propensities with sampling weights to find "highly influential" cases—but do not experimentally compare the effectiveness of targeting to the effectiveness of randomly provided incentives. Such a comparison is crucial, however, in evaluating the effectiveness of targeting.

The most similar approach to ours is Jackson, McPhee, and Lavrakas (2020), which estimates response propensities and uses these to target incentives to complete a screener for the National Household Education Survey (NHES).[57] As in our proposed design, Jackson, McPhee, and Lavrakas (2020) randomly divides potential respondents into a group that receives incentives independent of their propensity or one in which propensities determine incentive receipt. Specifically, the conditions are:

1. For the group assigned to propensity-independent incentives, respondents randomly receive either a $2 noncontingent incentive or a $5 noncontingent incentive along with their screener.

[56] The authors examine a different type of incentive—contingent or promised incentives—than the present study. With that in mind, they find no improvements in response rates or bias from a promise of $25 relative to $0, but a significant improvement in both response rates and bias from a promise of $45 compared to $25.

[57] The authors use a two-stage approach. First, they use a conditional inference tree for variable selection. Then, they use logistic regression with the selected variables.

2. For the group assigned to targeted incentives, low propensity cases received $10, medium propensity cases received $5, medium-high propensity cases received $2, and very high propensity cases received $0.

Jackson, McPhee, and Lavrakas (2020) represents an important step forward for research on targeted incentives. However, its design has a fundamental drawback: the only group in which respondents receive no incentives is the targeted group. Thus, the effect of targeting is confounded with the effect of receiving no incentives. Unsurprisingly, giving high-propensity respondents $0 (in group 2) versus $2 or $5 (in group 1) decreases the response rate substantially. Thus, the study does not provide a good test of the targeting mechanism per se because it confounds targeting with the lack of incentives. Furthermore, predicting response based on demographic variables alone is notoriously difficult. Because the AHS is a panel survey of housing units, we are able to take advantage of a richer set of longitudinal data to improve the quality of the predictive models. In particular, in addition to the sampling frame data, we are able to include prior wave response outcomes (i.e., whether or not the unit responded) and prior wave paradata, which include the number, type, and timing of contact attempts and reasons for refusing the survey. We additionally leverage time-varying neighborhood characteristics from respective American Community Survey (ACS) 5-year estimates (2014; 2016; 2018) that capture aggregate demographic characteristics (age; employment status) potentially related to nonresponse. These data sources lead to a high-dimensional dataset with 400+ predictors; we use machine learning classifiers to retain this high-dimensional predictor set in the predictions of nonresponse.[58]

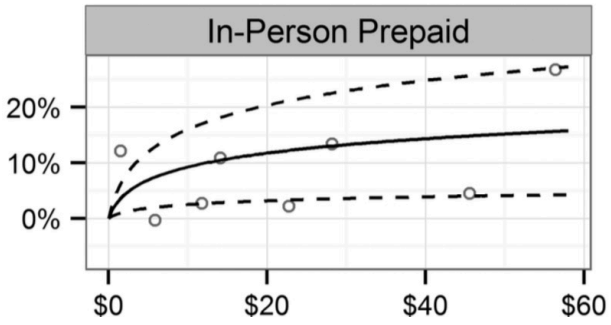
An additional point raised in Jackson, McPhee, and Lavrakas (2020) is that different incentive amounts may produce different kinds of responses as a function of predicted response propensities. However, because the varying incentive amounts are not randomized across different propensities, their study leaves this question largely unanswered.

2.3 What is the right incentive amount?

An early finding in the literature on incentives is that, while response rates increase as the incentive amount increases, they do so at a decreasing rate (Armstrong, 1975). In a large meta-analysis of the effect of incentive amounts on response rates, Mercer et al. (2015) showed that (1) the type of incentive and survey mode appeared to matter for the dose-response curve (see Figure 3 for their in-person dose-response curve) and (2) that a relative paucity of data on varying amounts in the context of mixed-mode, panel surveys such as the AHS made generalizing to those contexts based on extant literature difficult. Understanding where the inflection point lies in the AHS survey sample will help to determine whether a flat $5 incentive, as is used in the NHES, makes sense, or whether differing amounts need to be used among different subgroups.

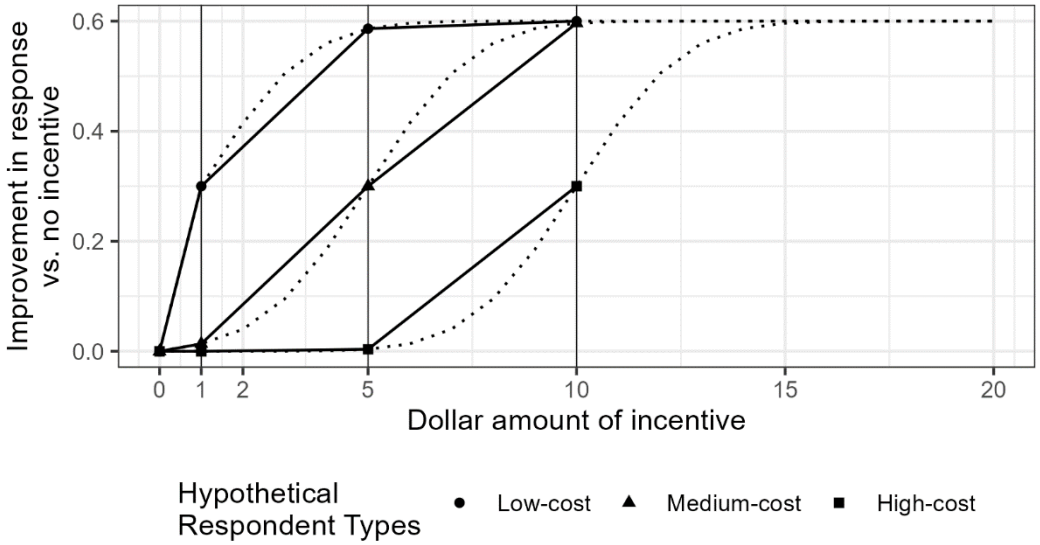
[58] This also contrasts with Jackson, McPhee, and Lavrakas (2020), who use a decision tree to reduce the dimensionality of the predictors and then a parametric logistic regression to generate predictions. In preliminary analyses, we find that more complex models significantly outperform decision trees.

Figure 3. Dose-Response Effect of Incentives on Response Rate for in-Person Surveys Using Noncontingent Incentives, Reproduced from the Mercer et al. (2015) Meta-Analysis



Our study plans to randomize respondents to one of four amounts: $0, $2, $5, and $10. The $5 dollar amount is chosen as it corresponds to amounts in similar surveys such as the NHES. Figure 4.3 demonstrates examples of the response curves we might find. The dotted curves illustrate unobservable dose-response curves, while the solid curves and points show estimable quantities that the design can elicit.

Figure 4. Possible Dose-Response Curves and Estimable Linear Relationships in the Proposed Study



We include the $2 amount as it is possible that we find ourselves in the low-cost, scenario, in which the bulk of the response rate increase can be generated with two dollars. However, the medium-cost scenario seems very plausible. Mercer et al. (2015), for example found that, on average, in person surveys that paid $5 versus nothing had a response rate increase of 5 percentage points, those that paid $10 versus nothing had an increase of 7 percentage points, while those that paid $20 had an increase of 9 percentage points. In other words, while doubling the incentive from 5 to 10 produced a 40 percent increase in effectiveness, doubling it from $10 to $20 only produced a 28 percent increase in effectiveness.

For this reason, we believe it makes sense to test an amount of $10. Moreover, the panel context of the AHS argues in favor of including at least one substantial incentive amount. In particular, it is important to know how incentives in one wave affect response patterns in subsequent waves. While respondents may very easily forget having received $2 or $5 two years ago given the

largely symbolic value of these sums, $10 seems more likely to stand out in one's memory. This raises the prospect that, either through habit formation or recall, large incentive amounts may durably increase response rates beyond the one wave in which they are conducted or lead to an expectation of similar incentives in future waves. This is a possibility largely unexplored in the literature.

2.4 Remaining Gaps in the Literature

While the literature we review below shows that incentives are effective at increasing response rates, there are three gaps, some of which the present study aims to fill but others that remain for future research.

First, despite recognition that increasing the response rate overall does not necessarily reduce nonresponse bias (Groves, 2006), studies largely continue to focus on response rates as the outcome to improve rather than measures of bias. In a meta-analysis published seven years after the points made by Groves (2006), Singer and Ye (2013) note an ongoing lack of research into the ability of incentives to address nonresponse bias. Since then, both Crissey, Christopher, and Socha (2015) and Jackson, McPhee, and Lavrakas (2020) target measures of bias as outcomes in addition to response rates, benchmarking characteristics of respondents to known quantities, but their studies have yielded no discernible improvements in bias from targeting.[59] Our study continues their work in examining reductions in bias, rather than improvements in response rates, as the primary outcome.

Second, Jackson, McPhee, and Lavrakas (2020) is the first study of which we are aware to experimentally compare the impact of (1) incentives given independent of a household's response propensity (respondents randomly assigned to $2 or $5) to (2) incentives based on response propensities, with higher amounts given to those with lower propensities and no incentive given to those with a high response propensity. However, because the second condition involved giving escalating incentives based on propensities, it does not allow us (1) to compare the full range of incentives ($0–$10) in all strata of response propensities or (2) to compare a strategy of randomly deciding who receives *any* incentive to a strategy of giving incentives to those with high nonresponse propensities. The design we outline below aims to fill these gaps.

Finally, and returning to the three groups we outlined above—(1) those with low response propensities; (2) those with low response propensities who have the highest likelihood of contributing to bias; (3) those with low response propensities, high bias-contribution likelihoods, and a high likelihood to be "moved" to respond by incentives—all existing research either targets group one (Jackson, McPhee, and Lavrakas, 2020) or a combination of groups one and two (Crissey, Christopher, and Socha, 2015; Coffey and Zotti, 2015). As Jackson, McPhee, and Lavrakas (2020) note:

> An important outstanding question is whether it is possible to classify cases based not only on their base response propensity but also on the increase in response propensity that would be attributable to (for example) a higher incentive. If cases are heterogeneous in their sensitivity to an intervention, and if this sensitivity can be predicted from auxiliary data available prior to collection, then it may be efficient to target the intervention based

[59] More precisely, Crissey, Christopher, and Socha (2015) only find improvements in bias when the promised incentive for completing the survey is $45.

on predicted sensitivity. (407)

Because our study will randomly allocate amounts across propensities, it will take a step toward addressing this gap. In particular, our study should permit the construction of "sensitivity scores" that will enable future incentive studies to test this third type of targeting.

2.5 Intervention Design

The intervention involves providing incentives randomly in one randomly selected half the sample and, in the other randomly selected half, providing incentives only to those predicted to not respond absent incentives. We define its features with the aid of some simple formal notation.

Let there be a universe, U, of individuals indexed i, who comprise a fixed and finite population of size N whose characteristics some decisionmaker would like to learn. Specifically, suppose that individuals have a feature, X_i, whose true mean the decisionmaker would like to learn: $\bar{X} = \frac{1}{N}\sum_{i \in U} X_i$. For example, this might represent the true rate of severely inadequate housing in the United States. To learn $\bar{X}$, the decisionmaker takes a random sample of n individuals. Let $S_i \in \{0,1\}$ denote a random variable that indicates selection into the sample. Sample probabilities are $\pi_i^S = \Pr(S_i = 1)$. We let $Y_i \in \{0,1\}$ denote an indicator for response, and R the set of individuals who are both sampled and who respond, $R = \{i: S_i = 1, Y_i = 1\}$. The decisionmaker can only observe the feature for those who are sampled and who respond. In order to learn about $\bar{X}$, she uses the weighted sample mean estimate $\hat{\bar{X}} = \sum_{i \in R} X_i w_i$, where w_i is a sampling or bias-adjustment weight that sums to 1 ($w_i = \frac{1/\pi_i^S}{\sum_{i \in R} 1/\pi_i^S}$).

Suppose that the decisionmaker has a fixed monetary budget, B, that she can use to incentivize potential respondents to respond to her survey. Denote by $b_i \in \mathbb{R}^+$, a positive dollar amount, the budget allocated to the i'th respondent. Suppose further that:

- the i'th potential respondent has an unobservable propensity to respond, $\eta_i = \Pr(Y_i = 1)$.
- η_i is correlated with the covariate of interest, X_i, which is either fixed and not changeable by attempts at contact (e.g., age) or measured prior to the attempt at contact (e.g., percentage of household income paid toward rent).
- propensities are increasing (monotonically but possibly nonlinearly) in b_i ($\partial \eta_i / \partial b_i > 0 \; \forall \; i$).
- $\eta_i \in (0,1) \; \forall \; i$.

The response rate for a given sample is given by $\bar{Y} = \frac{1}{n}\sum_{\{i:S_i=1\}}^{n} Y_i$. Since S_i is a random variable, we can define the expected response rate over random samples as $E[\bar{Y}]$. We can also define the expected sample mean of X_i over random samples as $E[\hat{\bar{X}}]$.

With this setup and sufficiently large samples (e.g., large enough n), the problem is that under a no-spending world ($b_i = 0 \; \forall \; i$), it follows that:

- some potential respondents will respond and others will not, so that the expected response rate is not 100 percent ($E[\bar{Y}] \neq 1$), which increases uncertainty by increasing the variance of the sample mean estimate ($E[\hat{\bar{X}}^2] - E[\hat{\bar{X}}]^2$).

- respondents will have different covariate profiles than nonrespondents, with nonresponse bias defined as $\bar{X}-E[\hat{\bar{X}}]$. In general, we expect covariates to differ between people who respond and those who do not (for example, responders may be older, on average, than nonresponders).

This situation represents the status quo, in which no incentives are used. In expectation, decisions made on the basis of some $\hat{\bar{X}}$ will be less certain as $E[\bar{Y}]$ decreases (lower expected response rate), and more biased as $\bar{X}-E[\hat{\bar{X}}]$ increases in absolute size. The problem is thus to improve decisionmaking by devising some optimal way of allocating incentives, $\mathbf{b}^*$ (with $0 \leq b_i^* \leq B \; \forall \; i$), so as to achieve **two aims**:

1. Maximize the expected response rate, $E[\bar{Y}]$.
2. Minimize nonresponse bias, $|\bar{X}-E[\hat{\bar{X}}]|$.

Informally, what might an optimal $\mathbf{b}^*$ look like? Focusing firstly on the response rate, it seems obvious that spreading the budget too thinly is unlikely to provoke any change in response: providing someone with five cents might not be enough. So, unless B is very large or propensity to respond is highly responsive to even very small increases in incentive amounts, the strategy in which every respondent is given an equal share of B (i.e. $b_i = B/n$) is dominated by one in which a subset of size $m < n$ of all potential respondents is provided a cash incentive. For example, if the expected response rate can be reliably calculated, one might set $m = (1 - E[\bar{Y}])n$, so that the proportion of the sample that receives incentives, m/n, is equal to the proportion expected to not respond.

As noted above, this raises the question of how much does the incentive needs to be concentrated in order to cause a substantial increase in response: e.g., are two dollars enough or are five dollars necessary? Are there diminishing marginal returns, such that, for example, providing two dollars versus no dollars increases response much more than providing twelve versus ten dollars (i.e., $\partial^2\eta_i/\partial^2 b_i < 0$)? Providing accurate answers to these questions ensures that neither too much nor too little is spent on incentives in order to achieve the two aims.

It also raises the question, addressed only imperfectly in the literature described above, of how that subset should be chosen. If it were possible to glean information on propensities to respond, would allocating incentives to those with the lowest η_i increase response?[60] In addition, if the targeting is to those with the lowest propensity to respond, is there a subset of these low-propensity individuals who are most likely to introduce bias if not incentivized—that is, individuals that have attributes of interest that differ from those with high response propensities? How large would the gains from such an approach be?

In practice, decisionmakers do not get to observe response propensities when deciding how to allocate incentives. Moreover, incentives are often used in the context of experiments. Thus, more often than not, incentives are allocated independently from any potential respondent characteristics.

[60] This does not strictly have to be the lowest η_i but can be those with relatively lower propensities, for example, those deemed close to the margin of responding; however, the general logic of targeting is the same even if the selected set of propensities for targeting is somewhat shifted.

In theory, however, one potentially more optimal $\mathbf{b}^*$ would allocate none of the budget to those respondents who will respond even in the absence of incentives, because it is inefficient to offer incentives to those who would respond without the additional inducement. Allocating the incentive budget to those most likely to contribute to nonresponse bias would instead optimize the incentive budget in line with the goals listed above.

Suppose that the decisionmaker has access to an estimated propensity, $\hat{\eta}_i$, which includes both the propensity to respond and a likelihood of introducing bias. There is a spectrum of ways in which she could allocate incentives to m respondents as a function of their estimated propensities. At one extreme of the spectrum, she might allocate incentives completely independently of propensities ($\Pr(b_i|\hat{\eta}_i) = \Pr(b_i)$). At the other end of the spectrum, she may allocate incentives to respondents as a deterministic function of their estimated propensity. We compare the two extremes of this spectrum of allocation mechanisms:

1. **Propensity-Independent Allocation**: incentives are allocated to potential respondents independently of their true or estimated propensities $\Pr(b_i|\eta_i) = \Pr(b_i)$.
2. **Propensity-Determined Allocation**: potential respondents are indexed in order of their estimated response propensities, $\hat{\eta}_i$, so that $\hat{\eta}_1 = \min(\hat{\eta}_i)$ and $\hat{\eta}_n = \max(\hat{\eta}_i)$. The key feature of this assignment is that incentives are deterministically provided to those respondents deemed most at risk of nonresponse ($\Pr(b_i > 0|i \leq m) = 1$) and no incentive is provided to the rest of the respondents ($\Pr(b_i > 0|i \geq m) = 0$). In addition, this propensity-determined allocation may compare (1) different methods for estimating the propensity (e.g., comparing a simple rule based on previous nonresponse behavior to a more complex model) and (2) may target modifiable forms of nonresponse (e.g., refusals in previous waves) rather than all forms of nonresponse.

3 Evaluation Design

We are interested in understanding how propensity-determined allocation of incentives affects nonresponse bias and how the size of incentives delivered to a potential respondent affects the rate of response among different subgroups in the sample. The randomization is designed to generate the counterfactuals necessary to make these quantities estimable. We define these counterfactuals below.

Continuing from the formalization above, the evaluation imagines that those sampled into the 2021 AHS survey could have been allocated incentives using either of the two allocation mechanisms above. Let $Z_i \in \{0,1\}$ denote a random variable that indicates whether potential respondent i has been assigned to receive an incentive.

First, we denote by $Z^{T=0}$ the allocation of incentives that would have obtained had **Propensity-Independent Allocation** been used for the entire sample. An n-length vector of m 1s and $n - m$ 0s is generated, in which there is no dependence of the assignment on estimated propensities: $\Pr(Z_i^{T=0} = 1|\hat{\eta}_i) = Pr(Z_i^{T=0} = 1) \approx .3$. The 1s and 0s are simply shuffled among the potential respondents.

Second, we denote by $Z^{T=1}$ the allocation of incentives that would have obtained had **Propensity-Determined Allocation** been used for the entire sample. The potential respondents are sorted by their $\hat{\eta}_i$, from lowest to highest, and an n-length vector of m 1s and $n - m$ 0s is generated (with the 1s at the top and the 0s at the bottom.) Thus, those m respondents with the lowest 30 percent of estimated propensities are guaranteed to receive an incentive, and those $n -$

m respondents with the highest 70 percent of estimated propensities are guaranteed not to receive an incentive. Note that this represents a considerable advantage: with an expected response rate of 74 percent, if our model does well at predicting nonresponse, we would be targeting all predicted nonrespondents—including both those on the margin and those who are perhaps less likely to be converted to responses—but a minimum of those already likely to respond.

We define the vectors $Z^{T=0}$ and $Z^{T=1}$ for the full sample: these are the assignments that *would* obtain, *were* we to use propensity-independent or -determined allocation methods for the full survey. These are the allocation counterfactuals.

From here, we suppose that every respondent has a potential outcome function, $Y_i(Z_i)$. In particular, we imagine that $Y_i(Z_i^{T=0} = 1) = Y_i(Z_i^{T=1} = 1)$ and $Y_i(Z_i^{T=0} = 0) = Y_i(Z_i^{T=1} = 0)$, so that if the potential respondent would have (not) responded when (not) assigned to an incentive under one allocation scheme, they also would have (not) responded when (not) assigned to an incentive under the other. Some research has shown that knowing that one is in a lottery-style incentive condition versus deterministic condition could matter for responses. However, since respondents will not know that they are being randomly assigned to conditions here, we don't have reason to doubt this assumption.

This stability in the potential outcomes allows us to define, for a given $Z^{T=0}$ and $Z^{T=1}$, the outcomes that would have resulted had one or the other allocation schemes been used to assign incentives.

The experiment works by generating $T_i \in \{0,1\}$ (for "targeting"): when $T_i = 0$, the individual is given the Z_i corresponding to $Z^{T=0}$ and they reveal the Y_i that corresponds to $Y_i(Z_i^{T=0})$; when they are given $T_i = 1$, they are given the value of Z_i that corresponds to $Z_i^{T=1}$, and reveal the outcome that corresponds to $Y_i(Z_i^{T=1})$. The targeting variable, T_i, is generated by sorting individuals by an estimated propensity to respond, forming consecutive pairs, and flipping a virtual coin within each pair. We thereby obtain one "random sample" from the world in which we did propensity-determined allocation and one from the world in which we did propensity-independent allocation. The pairs ensure that, for any given tranche of propensities, there will be near-perfect balance with respect to T.

One concern with such a procedure is that it generates correlation between Z_i and $\hat{\eta}_i$ and X_i. In other words, the assignment creates confounding between propensity to respond, probability of assignment to treatment, and the characteristics we care about.

As it turns out, however, this is a simple case of heterogeneous assignment probabilities. And, as we show below, it is easily dealt with using an inverse propensity weighted estimator. Specifically, since T is independent, for any given individual the probability of assignment is given by $\Pr(Z_i = 1) = \Pr(T_i = 1) \times Pr(Z_i = 1|T_i = 1) + \Pr(T_i = 0) \times Pr(Z_i = 1|T_i = 0)$. For the 30 percent (m/n) of units with the lowest propensity to respond (who will be allocated an incentive under targeting), this evaluates to $.5 \times 1 + .5 \times .3 = .65$. For the 70 percent of units with the highest propensity to respond (who will not be allocated an incentive under targeting), this evaluates to $.5 \times 0 + .5 \times .3 = .15$. Thus, there are four possible values of a treatment assignment probability $\pi_{i,z}^{Z}$ (where z indicates an *observed* treatment status): for j low propensity individuals, $\pi_{j,1}^{Z} = .65$ and $\pi_{j,0}^{Z} = 1 - .65 = .35$; for k high propensity individuals, $\pi_{k,1}^{Z} = .15$ and $\pi_{k,0}^{Z} = 1 - .15 = .85$. Thus, it is possible to observe every unit in every treatment condition,

albeit with differing probabilities. To obtain unbiased estimates of the average treatment effect, we simply downweight those who are overrepresented in treatment or control, and upweight those who are underrepresented, using $1/\pi_{i,z}^{Z}$, the inverse treatment propensity.

Note that there is no biasing path that confounds T and other outcomes of interest, such as Y_i or $\hat{\hat{X}}$. This drastically simplifies the estimation of unobservable quantities such as the proportion of respondents with $X_i = 1$ who would respond to the survey, if a propensity-determined allocation method were used for the whole sample: $E[\hat{\hat{X}}|T_i = 1]$. In simulation studies, we are thus well positioned to see both whether the deterministic allocation would produces an *actual* increase in the representativeness of the sample, and also whether our estimators are able to recover this.

Finally, while the variation in T that generates variation in Z is the main variation we are interested in, we are also interested in the elasticity of incentives to response: $\partial\eta_i/\partial b_i$ and $\partial^2\eta_i/\partial^2 b_i$. Thus, among those m assigned to incentives, we plan to vary the amount of the incentive between 2, 5, or 10 dollars. This enables us to study the change in η_i induced by a one-unit change in b_i. As we describe in greater detail below, this dose-response function could be highly nonlinear. However, we are able to recover an estimand that is defined as a linear transformation of the potential outcomes using a linear estimator, even though the potential outcomes are generated through a nonlinear process. See Figure 4.3 above for a graphical illustration of this point.

3.1 Total Number of Observations

The 2021 AHS integrated national sample will build on the existing panel created by sampling just over 85,000 units in 2015. We anticipate that the final sample will be close to 84,000.

3.2 Randomization and Assignment

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:

1. **Create $\boldsymbol{Z_i^{T=1}}$**. 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 $\boldsymbol{Z_i^{T=0}}$**. 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 $\boldsymbol{T_i}$**. Sort individuals in order of their estimated propensity (randomly resorting 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 percent to $A_i = 10$, 25 percent to $A_i = 5$, and 25 percent to $A_i = 2$. Assign the remaining sample for whom $Z_i = 0$ to $A_i = 0$.

3.3 Treatment Conditions

The random assignment of the three variables, A, Z, and T, results in eight treatment conditions. The large number of conditions may sound like it puts the study at a risk of low power, but in practice the study is not analyzed as a multi-arm design. Mostly, estimands are defined by marginalizing over conditions to obtain a difference in two conditions. The table below translates the procedure above into proportions and sample sizes, based on an approximate sample size of 84,000.

3.4 Outcomes

	Propensity-Independent (50%)				Propensity-Determined (50%)			
Incentive $ amount	0	2	5	10	0	2	5	10
Incentive proportion	70%	7.50%	7.50%	15%	70%	7.50%	7.50%	15%
Total number	29,400	3,150	3,150	6,300	29,400	3,150	3,150	6,300
Sample proportion	35%	3.75%	3.75%	7.50%	35%	3.75%	3.75%	7.50%

At this stage, we are interested in three main outcomes, and three secondary outcomes. These will likely evolve somewhat as we begin to refine the analysis plan.

Main Outcome: Effect of propensity-determined allocation on the difference in sample and population mean of key outcome or covariate

- Interpretation: This outcome focuses on whether propensity-determined incentive allocation makes sample estimates of outcomes such as home ownership less biased and, when concerning demographic variables, whether it improves representativeness. We discuss measures of representativeness in the AHS nonresponse bias memo Sections 2 and 5. The main outcome is the distance of the mean of X_i in the sample versus in some reference population. For example, X_i may be a binary indicator for whether the householder owns the housing unit outright, which is a key outcome for the AHS that is also measured in the Decennial Census. Our separate analysis suggests this quantity is overestimated, even when using bias adjustment weights, so that $\bar{X} - E[\hat{\bar{X}}]$ will be strictly positive. We expect that changing from random to deterministic allocation decreases this quantity.
- Definition of estimand: Denoting $\bar{X}$ the true population mean of X_i, and $E[\hat{\bar{X}}|T = t]$ the estimated mean of X_i among those in the sample who respond when the allocation mechanism is t, our estimand is: $(\bar{X} - E[\hat{\bar{X}}|T = 1]) - \bar{X} - E[\hat{\bar{X}}|T = 0])$.
- How we estimate it: Regress the distance of $D_i = \bar{X} - X_i$ on T_i. We refer to this estimand as "Effect of T on sample vs pop. Mean(X)" in design diagnosis below.

Main Outcome: Effect of propensity-determined allocation on response rate

- Interpretation: This is the average effect of propensity-determined allocation on the overall response rate. Per the formalization above, we should expect propensity-determined allocation to increase the overall response rate relative to propensity-independent allocation, as well as increasing representativeness.
- Definition of estimand: $\frac{1}{n}\sum_{\{i:S_i=1\}} Y_i(T=1) - Y_i(T=0)$.
- How we estimate it: We regress Y_i on T_i. In the design diagnosis below, we refer to this estimand as the "Effect of T on Y".

Main Outcome: Effect of a one-dollar change in incentive amount on response rate

- Interpretation: This outcome measures how much a one-dollar change in the amount of the incentive increases average response rates, linearly. Our estimand is a parameter from a model applied to the potential outcomes: it can be thought of as the coefficient one would get on A if one were to able to fit a least squares model to all possible potential outcomes on all possible conditions for all units. Note: we are thinking of A as continuous under this definition.
- Definition of estimand: the β that solves:

$$\min_{(\alpha,\beta)} \sum_i \int (Y_i(x) - \alpha - \beta A)^2 f(A) dA$$

- How we estimate it: We regress Y_i on A_i in a weighted least squares model, in which the weights are the inverse of the probability of observing unit i in condition $A_i = a$. In other words, each unit's contribution to the likelihood is weighted by $\frac{1}{\Pr(A_i=a)}$. In the design diagnosis below, we refer to this estimand as the "Change in Y caused by unit change in A."

Main Outcome: Effect of being sent an incentive on response rate

- Interpretation: This is the average effect of being sent any incentive on the response rate.
- Definition of estimand: Assuming homogeneous effects for incentive amounts for ease of exposition, it is simply $\frac{1}{n}\sum_{\{i:S_i=1\}} Y_i(Z=1) - Y_i(Z=0)$. Under heterogeneous effects, it is the average of the unit-level averages of three estimands: $\frac{1}{n}\sum_{\{i:S_i=1\}} Y_i(A=10) - Y_i(A=0)$, $\frac{1}{n}\sum_{\{i:S_i=1\}} Y_i(A=5) - Y_i(A=0)$, and $\frac{1}{n}\sum_{\{i:S_i=1\}} Y_i(A=1) - Y_i(A=0)$.
- How we estimate it: We regress Y_i on Z_i in a weighted least squares model, in which the weights are the inverse of the probability of observing unit i in condition $Z_i = z$. In other words, each unit's contribution to the likelihood is weighted by $\frac{1}{\Pr(Z_i=z)}$. In the design diagnosis below, we refer to this estimand as the "Effect of A>0 on Y."

Secondary Outcome: Effect of propensity-determined allocation on sample mean of covariate

- Interpretation: This is the average effect of propensity-determined allocation on the mean of some covariate. This can be thought of as a more direct estimate of bias in the sense

that we are able to directly observe estimates of key outcomes of interest for both treatment conditions. If propensity-determined allocation changes the proportion of groups likely to introduce bias above a propensity-independent allocation, we should be able to estimate this increase.

- Definition of estimand: $E[\hat{\hat{X}}|T_i = 1] - E[\hat{\hat{X}}|T_i = 0]$.
- How we estimate it: We regress X_i on T_i. In the design diagnosis below, we refer to this estimand as the "Effect of T on sample mean(X)."

Secondary Outcome: Effect of incentives on number of contact attempts

- Interpretation: This is the average effect of being sent any incentive on the number of contacts attempted with a respondent (successful and unsuccessful interviews).
- Definition of estimand: $\frac{1}{n}\sum_{\{i:S_i=1\}} Y_i(Z = 1) - Y_i(Z = 0)$.
- How we estimate it: We regress Y_i on Z_i in a weighted least squares model, in which the weights are the inverse of the probability of observing unit i in condition $Z_i = z$.

3.5 Meaningful Effect Size

In our simulation studies of his design, we define potential outcomes in the following way:

$$Y_i(Z_i = 0) = \text{Binom}(\eta_i) \tag{1}$$

$$Y_i(Z_i = 1) = \begin{cases} 1 & \text{if } Y_i(Z_i = 0) = 1 \\ \text{Binom}(\tau) & \text{if } Y_i(Z_i = 0) = 0. \end{cases} \tag{2}$$

Where τ is the effect of the incentive on if-untreated nonresponders (those for whom $Y_i(Z_i = 0) = 0$). Providing incentives is assumed here to only affect those who would not have responded when no incentive was provided, and only increases the likelihood of response: we rule out cases where providing an incentive causes nonresponse in someone who would have responded in the absence of incentives (although, in theory, such cases are possible – say, if control responders are so offended by receiving a dollar they decide not to respond).

Thus, we can distinguish between τ, the average effect of receiving an incentive among if-untreated nonresponders, and $\overline{\tau}$, the average effect of receiving an incentive in the sample.

We think that anything above a 1 percentage point increase in the overall response rate is a meaningful effect. Note that $\overline{\tau} = (1 - \bar{Y}(0))\tau$. In the 2017 AHS, for which we can only observe units in the control condition, we have $\bar{Y}(0) \approx .80$. So, we can back out the (constant) effect incentives would have to generate among if-untreated nonresponders in order to obtain $\overline{\tau} = .01$ using $.01 = (1 - .80)\tau$, which implies $\tau = .01/.20 = .05$. Thus, in order to observe a sample average treatment effect of a 1-percentage point increase in the response rate ($\hat{\tau} = .01$), incentives would need to increase the response probability of if-untreated nonresponders by five percentage points on average. This seems like a reasonable bar to clear.

3.6 Likely Effect Size

Singer et al. (1999) compared the results of 39 experiments on financial incentives in face-to-face and telephone surveys. The effect sizes were smaller (though not statistically significantly

on average, implying a likely average effect of $.003 \times 6.75 = .02 = \overline{\tau}$. In terms of average effects on if-untreated nonresponders, this implies a 10 percentage point increase ($\tau = .10$). These parameters are assumed in the power calculations below.

3.7 Power

Using DeclareDesign, we conducted a preliminary diagnosis of the design's ability to estimate the outcomes described above, assuming $\tau = .10$ and $\overline{\tau} = .02$. In addition to power, we are able to diagnose the bias, coverage, and variance properties of the different estimator-estimand pairs.

Estimand Description	Mean Estimate	Mean Estimand	Bias	Power	Coverage	SD Estimate	Mean SE
Effect of T on sample vs pop. Mean (X)	– 1.20	– 1.20	0.00	1.00	0.97	0.10	0.11
Change in Y caused by unit change in A	0.70	0.70	0.00	1.00	0.96	0.04	0.04
Effect of A > 0 on Y	2.00	1.93	0.07	1.00	0.95	0.29	0.31
Effect of T on Y	0.98	1.00	– 0.01	0.95	0.96	0.27	0.27
Effect of T on sample mean (X)	1.20	1.20	– 0.00	1.00	0.97	0.10	0.11

The numbers are scaled to reflect percentage point changes. The first row can thus be interpreted as follows: the average estimate of the "Effect of propensity-determined allocation on the difference in sample and population mean of covariate" is -1.20 percentage points, and so is the average value of the estimand. Thus, the bias for this estimator is zero. The power is 1, implying the design is able to reject the null given the true underlying -1.20 percentage point reduction. The 95 percent confidence interval covers the true estimand 97 percent of the time. This is most likely indicative of simulation error, and possibly some slight conservative bias in the standard errors, which is to be expected. The standard deviation of estimates across the sampling distribution generated by the simulations – the "true" standard error – is one-tenth of a percentage point, which is approximately equal to the average standard error estimated (again, the standard errors appear very slightly conservative). Overall, the average estimate is ten times greater than the average standard error, indicating a high degree of statistical power. The conclusion is that the design does a very good job of estimating an increase in representativeness using this particular definition of representativeness (decrease in underrepresentation of $X = 1$).

Moving to the rest of the table, the estimators and estimands are all signed as we would expect. Proceeding row by row: the propensity-determined allocation method produces less distance in estimates of x compared to the propensity-independent method; each extra dollar has a linear effect on the response rate equal to .70 percentage points; receipt of any incentive increases the response rate by two percentage points on average; propensity-determined allocation increases the response rate by one percentage point more on average than propensity-independent allocation does; and propensity-determined allocation also increases the proportion of respondents with $X_i = 1$ (the simulations assume such respondents are ordinarily underrepresented).

In general, the estimators are all well powered given the large sample size and the assumptions of the simulation. Comparing point estimates to standard errors, the change in Y caused by unit change in A estimator is clearly the most efficient: the point estimate is over seventeen times larger than the standard error on average. This estimator is thus our best powered.

There is a very small amount of bias in two of the estimators. This is likely due to simulation error, either in the simulations for the diagnosis, or in the simulations used to generate assignment weights. It is small enough, at less than one-tenth of a percentage point, as to be negligible. As mentioned, there is little concern for false positives from the standard errors: if anything, they exhibit a small amount of the well-documented Neyman standard error bias that results from underestimation of the covariance in potential outcomes.

3.8 Data

We currently have access to the 2015, 2017, and 2019 AHS Integrated National Samples. We also have datasets we will use to estimate propensities, namely: the 2018 public-access Census Planning Database, as well as the (1) AHS 2015, 2017, and 2019 "CHI" datasets, which provide paradata on nonresponse for all units, and (2) trace files for all three waves that provide more detail on each unit's progression through the survey instrument. These data are sufficient to conduct randomization and hand off to partners at the Census Bureau.

3.9 Anticipated Limitations

There are a handful of risks worth highlighting. For the first three, we have conducted analyses that we outline in the accompanying summary memo–"Nonresponse Bias in the American Housing Survey 2015-2019"–that address the first three limitations.

1. **Our propensity model may not be good**. The design assumes that we are able to estimate η_i in a reasonably informative way. If we don't have good propensity estimates, then any allocation of incentives on the basis of such estimates will be weaker. However, we are in a very favorable context in this study: we have panel data that has two years' worth of information about how respondents behaved in the past, as well as tract-level demographic information from the American Community Survey.
    - *How we address*: In section 3 of the summary memo, we show that we can predict both nonresponse and refusal with a very high degree of accuracy in the 2017 and 2019 AHS. The most important predictors are past behavior—e.g., if a unit was a refuser in 2017 they are significantly more likely to continue to refuse in 2019. However, area-level demographics were also important predictors. We will use these findings to improve our ability to estimate η_i in the targeting experiment.
2. **Developing estimates of X from AHS data will be complicated**. The AHS data are not a simple random sample—data need to be reweighted to account for the sampling procedure in order to generate estimates. And our data also need to be weighted by the inverse of the assignment propensities. So there is some complication here that is something of a risk—we need to make sure we get the weights right in order to say something meaningful about representativeness.
    - *How we address*: in the summary memo, we discuss two considerations when comparing the AHS to benchmark data (in that case, the Decennial census). First is making sure that we align the variable definitions in each of the samples, which includes ensuring that we compare households to other households and that we compare questions asked in similar ways. Second is reweighting to account for the complex survey design. In the memo, and in follow-up discussions on proper weighting methods, we believe we can generate estimates of X to properly compare to a benchmark population, both at the national and the CBSA level.

3. **Response bias may not be strong enough to detect a reduction**. If the magnitude of nonresponse bias is small, any correction of them will be very small, and thus hard to detect. There is not a great deal we can do about this risk – we have designed as well powered a study as we can. We could possibly think about how to include covariates or focus the estimation on areas where underrepresentation is particularly strong.

    - *How we address*: the memo indicates substantial divergence between the AHS and the benchmark for certain characteristics. We believe this divergence is large enough to leave room for reductions in this distance.

4. **Spillovers due to stopping rule**. The Census Bureau typically stops data collection once the target of an 80 percent response rate has been met. This poses a spillover concern for us: 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 percent response rate.

    - *How we address*: The spillover issue is of particular concern for allocation methods that target at the area level. To address this, we have located our randomization at the respondent-level, where shifts in allocation of effort, which are coordinated by field officers, are unlikely. To assess *robustness* to spillovers, we will specify in the analysis plan a stop date, before which we believe spillovers of this kind will have kicked in, and at which we will estimate effects.

References

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review* 99 (1): 544–555. https://doi.org/10.1257/aer.99.1.544.

Armstrong, J.S. 1975. "Monetary Incentives in Mail Surveys," *Public Opinion Quarterly* 39 (1): 111–a16.

Coffey, S, and A. Zotti. 2015. "Implementing Static Adaptive Design in the National Survey of College Graduates Using the Results of an Incentive Timing Experiment." In *Joint Statistical Meetings*.

Crissey, Sarah, Elise Christopher, and Ted Socha. 2015. "Adaptive Design Strategies for Addressing Nonresponse Error in NCES Longitudinal Surveys," 28.

Edwards, Phil, Ian Roberts, Mike Clarke, Carolyn DiGuiseppi, Sarah Pratap, Reinhard Wentz, and Irene Kwan. 2002. "Increasing Response Rates to Postal Questionnaires: Systematic Review," *BMJ* 324 (7347): 1183. https://doi.org/10.1136/bmj.324.7347.1183.

Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys," *Public Opinion Quarterly* 70 (5): 646–675. https://doi.org/10.1093/poq/nfl033.

Groves, Robert M., Eleanor Singer, and Amy Corning. 2000. "Leverage-Saliency Theory of Survey Participation: Description and an Illustration," *The Public Opinion Quarterly* 64 (3): 299–308. https://www.jstor.org/stable/3078721.

Hidi, Suzanne, and K. Ann Renninger. 2006. "The Four-Phase Model of Interest Development," *Educational Psychologist* 41 (2): 111–127. https://doi.org/10.1207/s15326985ep4102_4.

Jackson, Michael T., Cameron B. McPhee, and Paul J. Lavrakas. 2020. "Using Response Propensity Modeling to Allocate Noncontingent Incentives in an Address-Based Sample: Evidence from a National Experiment," *Journal of Survey Statistics and Methodology* 8 (2): 385–411. https://doi.org/10.1093/jssam/smz007.

Laurie, Heather, and Peter Lynn. 2008. "The Use of Respondent Incentives on Longitudinal Surveys." 2008-42. Institute for Social; Economic Research. https://ideas.repec.org/p/ese/iserwp/2008-42.html.

Link, Michael W., and Anh Thu Burks. 2013. "Leveraging Auxiliary Data, Differential Incentives, and Survey Mode to Target Hard-to-Reach Groups in an Address-Based Sample Design," *Public Opinion Quarterly* 77 (3): 696–713. https://doi.org/10.1093/poq/nft018.

Mercer, Andrew, Andrew Caporaso, David Cantor, and Reanne Townsend. 2015. "How Much Gets You How Much? Monetary Incentives and Response Rates in Household Surveys," *Public Opinion Quarterly* 79 (1): 105–129. https://doi.org/10.1093/poq/nfu059.

Singer, Eleanor, John Van Hoewyk, Nancy Gebler, Trivellore Raghunathan, and Katherine McGonagle. 1999. "The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys," 14.

Singer, Eleanor, and Cong Ye. 2013. "The Use and Effects of Incentives in Surveys," *The ANNALS of the American Academy of Political and Social Science* 645 (1): 112–141. https://doi.org/10.1177/0002716212458082.

Appendix C: 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 overreliance 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,[61] 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

[61] The AHS raking procedure, as implemented in the 2019 wave, is described in Section 3.4 (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.

(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 during 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 = 43,000 (50%)				N = 43,000 (50%)			
Random assignment of **A** (dollar amount received)							
$0	$2	$5	$10	$0	$2	$5	$10
30,000	3,200	3,200	6,500	30,000	3,200	3,200	6,500
70%	7.5%	7.5%	15%	70%	7.5%	7.5%	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 nonadjusted 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:

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.

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.

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.[62]

[62] 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

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

- 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.[63] For this reason, all point estimates will be weighted (options two and three).
- 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.
- 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.[64] Otherwise, none of the data are 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.

---

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).

[63] 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" (U.S. Census Bureau and Department of Housing and Urban Development 2020: 3).

[64] These will be available in March 2022.

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 and 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 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:

$$Pr(Z_i = 1) = Pr(T_i = 1)Pr(Z_i = 1|T_i = 1) + Pr(T_i = 0)Pr(Z_i = 1|T_i = 0).$$

For the 30 percent (m/n) of units with the lowest propensity to respond,[65] (who are allocated an incentive under targeting), this evaluates to $0.5 \times 1 + 0.5 \times 0.3 = 0.65$. For the 70 percent 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 $\pi_{i,z}^{Z}$ (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): $\pi_{j,1}^{Z}$ = 0.65.
    - Assigned to control (no incentive): $\pi_{j,0}^{Z}$ = 1-0.65 = 0.35.
2. For k high propensity to respond (low propensity to nonrespond) individuals:
    - Assigned to treatment (any incentive): $\pi_{k,1}^{Z}$ = 0.15.
    - Assigned to control (no incentive): $\pi_{k,0}^{Z}$ = 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 $\frac{1}{\pi_{i,z}^{Z}}$, the inverse propensity weight (IPW).

[65] Or conversely, the highest propensity to not respond.

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:[66]

Own house (no; yes with mortgage/loan; yes with no mortage/loan).
Average household size.
White alone (householder).
Age of householder.
Rodents.
Mold.
Sampling frame: Census Division.
Sampling frame: HUD-assisted unit (as of 2013).
Sampling frame: 2013 Metropolitan Area (county-level; principal city, nonprincipal city, micropolitan area, non-CBSA area).
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 $T = 0$.

[66] 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.

- **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 percent of the control group owns their homes, while 63 percent of the treatment group owns their homes, we may contextualize these differences with reference to the national homeownership rate measured in the Census (~65 percent). 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 percent 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.[67] 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 (~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:

- 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.
- Contact attempts: we define this outcome as a continuous variable based on the CHI data and aggregating contact attempts across all modes.[68]

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'

[67] 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.

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

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:

$T = 1$ **increases response rate and leads to different sample composition:** the treatment increased response rate and may have reduced nonresponse bias.

$T = 1$ **increases response rate but does not lead to different sample composition:** the treatment increased response rate but had no detectable impact on nonresponse bias.

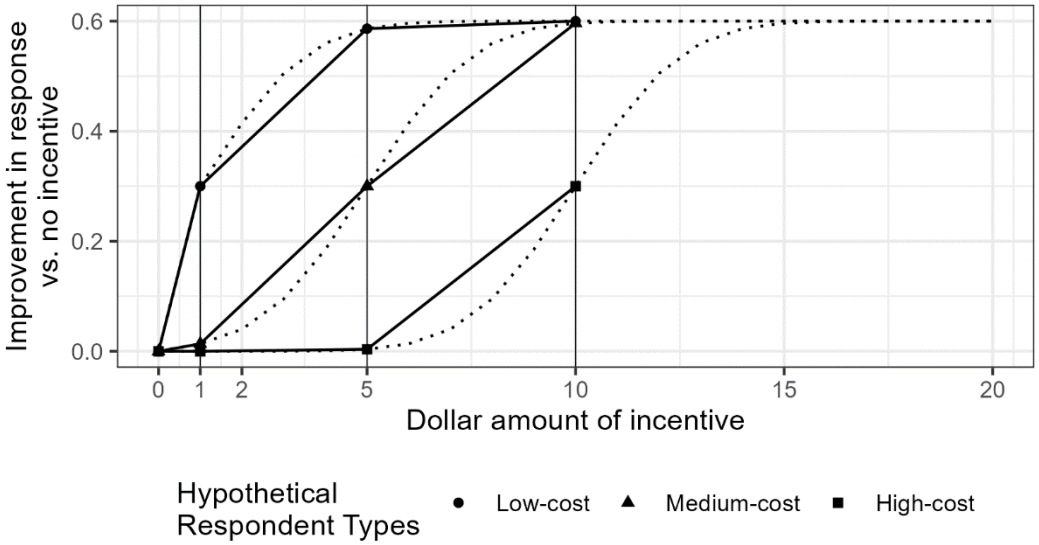
$T = 1$ **does not increase response rate but does lead to different sample composition:** the treatment changed the sample composition despite not changing the response rate.

$T = 1$ **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):

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: $3 \times 2 - 3 \times 0 = 5 \times 2 - 2 \times 2$.

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 ≥ \$5 to \$2 ≥ \$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:

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

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:

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

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 *pretreatment* 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 pretreatment 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 pretreatment 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.[69] 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).[70]

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 × two shifts in incentives).

4.11 Robustness Checks

We plan to conduct two robustness checks.

First is reestimating 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 percent 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 percent 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 percent (if relevant).
2. Reestimates 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 pretreatment attributes used in the AHS sample selection process (e.g., HUD-assisted as of 2013 or not) and pretreatment 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

[69] We may also use ACS contextual data.

[70] As described in Künzel, Sekhon, et al. (2019), metalearners for the CATE involve two steps. First, the data are 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.

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 premultiply 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 percent training set (with five folds then used for cross-validation to tune hyperparameters) and (2) 20 percent 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 percent held-out set using the metrics we discuss in Section 5.1.6.

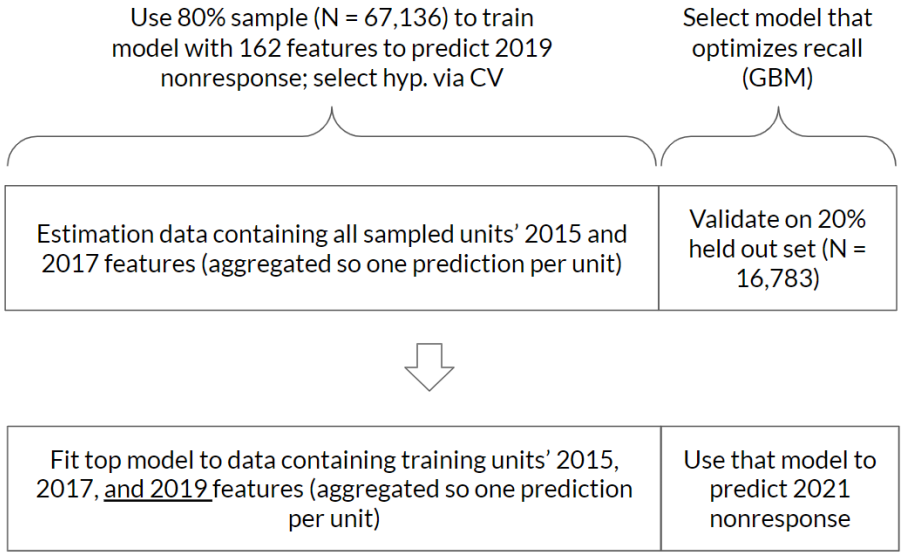
6. Finally, for all units and (1) using the best performing model, and (2) an updated feature matrix to which the 2019 values are added/aggregated, predict nonresponse in 2021 to generate the $\hat{\eta}_i$ used in the field experiment.

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.

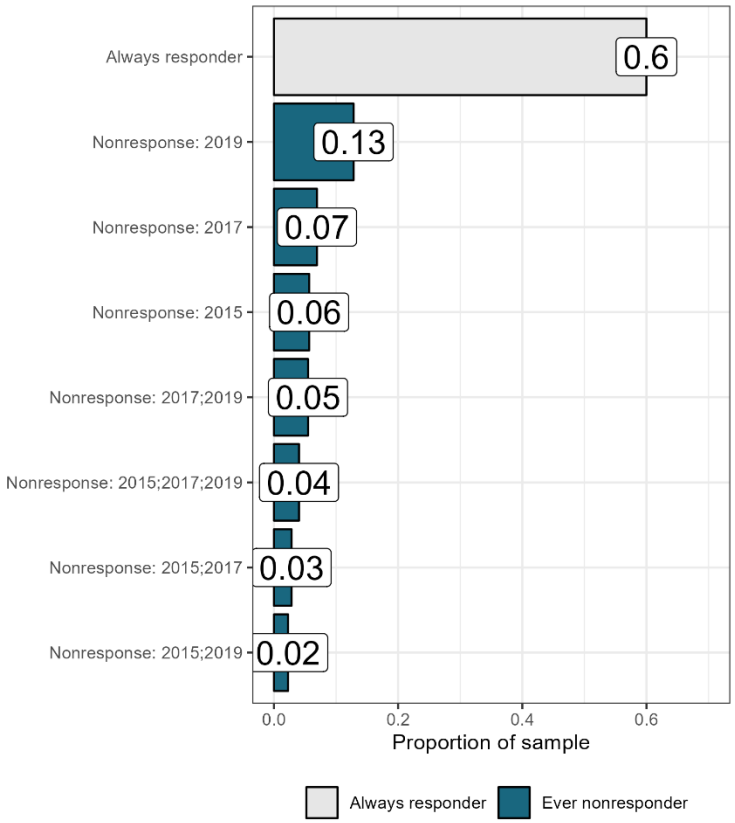
5.1.4 Baseline Predictions

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 (~25 percent 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 percent of units

are never responders, and over 10 percent 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. 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 percent 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.[71]

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

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

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

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. ŷ continuous	Pred. ŷ binary	True y	error_category
1537	0.99	NR	NR	Truepos.
1177	0.93	NR	NR	Truepos.
1879	0.84	NR	NR	Truepos.
1005	0.78	NR	R	Falsepos.
1187	0.72	R	R	Trueneg.
1034	0.71	R	R	Trueneg.
1159	0.60	R	NR	Falseneg.
1181	0.52	R	NR	Falseneg.
1071	0.49	R	R	Trueneg.
1082	0.47	R	R	Trueneg.
1603	0.44	R	R	Trueneg.
1762	0.33	R	R	Trueneg.
1319	0.29	R	R	Trueneg.
1359	0.24	R	NR	Falseneg.
1238	0.21	R	R	Trueneg.
1490	0.17	R	R	Trueneg.
1465	0.17	R	R	Trueneg.
1338	0.11	R	R	Trueneg.
1766	0.07	R	R	Trueneg.
1807	0.04	R	R	Trueneg.

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[72] and (2) baseline measures that correspond to the *status quo* methods survey planners might use.

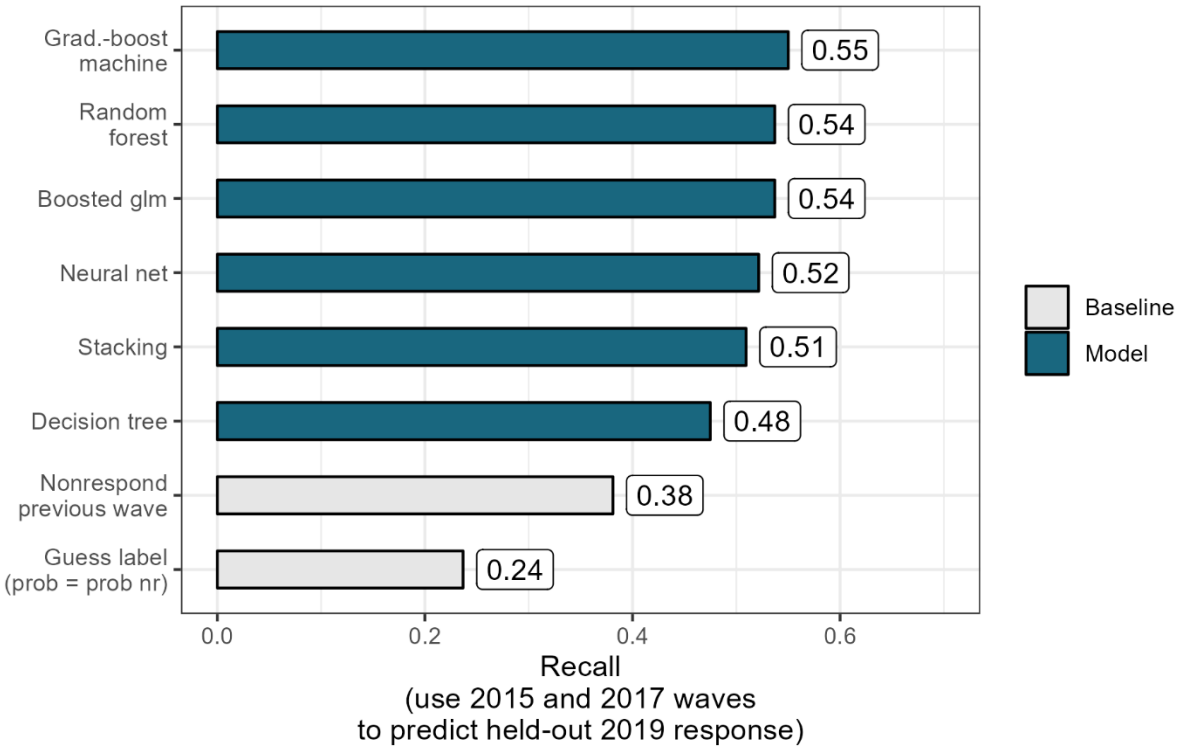
The graph shows that large gains in prediction come from the move from (1) *no targeting* (or just guessing nonresponse status based on its empirical proportion) to (2) even simple, rule-based targeting of using the nonresponse status in the previous wave to assume persistence in that behavior. The least 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

[72] 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.

nonresponse status, still substantially outperforms that simple rule (a gain in recall of 10 percentage points, or a 26.3 percent 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 percent improvement over that rule-based baseline.

Figure A3. Comparative Accuracy in 20 Percent 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) and (2) Risk-Based Targeting (Either Using a Single Model or Ensemble Method)

The figure shows recall metrics in the 20 percent 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 2 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.

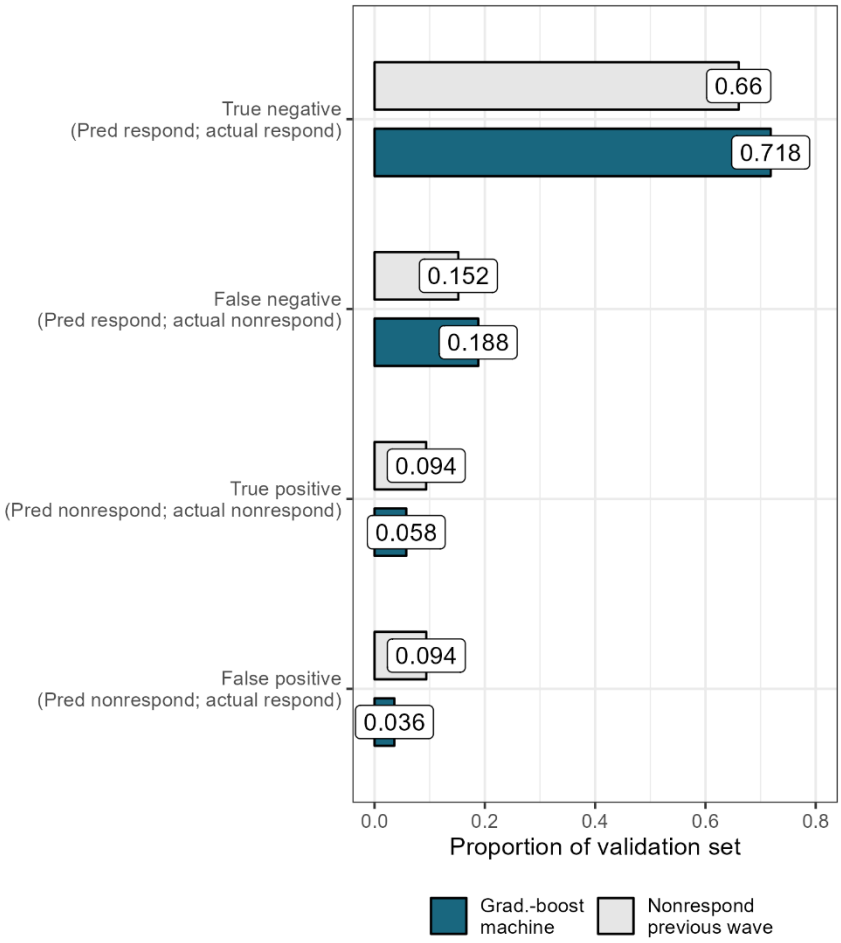
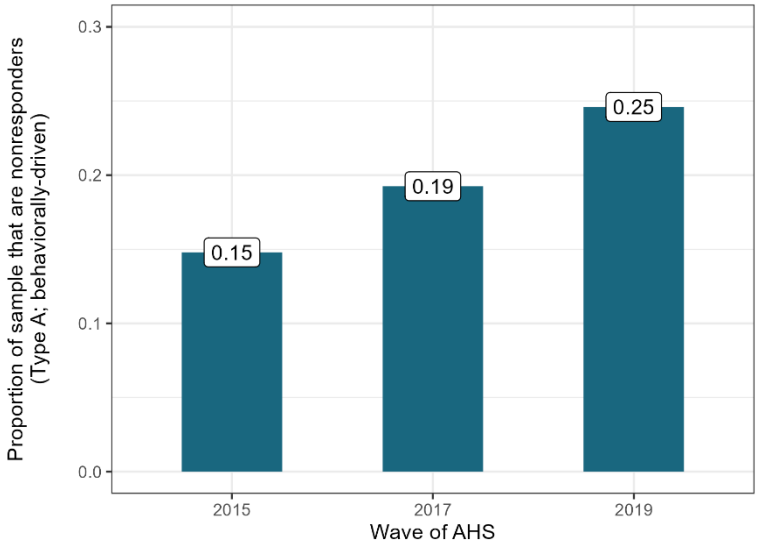


Figure A5. Possible Source of False Negatives in Model—Rising Base Rate of Nonresponse
The figure, again focusing on Type A, behaviorally driven nonresponse, shows how the leap in base rates may drive false negatives, or model-predicted responders who go on to nonrespond.



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 resorting 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 percent to $A_i = 10$, 25 percent to $A_i = 5$, and 25 percent 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 6.9.1, 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 ahs2019meth. 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.
    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:

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

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* 23 (2): 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.

Appendix D: Treatment Letters

Control: No Cash

AHS-26/66(L) (Los Angeles)

UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001
OFFICE OF THE DIRECTOR

Thank you in advance for your participation in this survey.

Dear Resident,

Your address has been selected by the U.S. Census Bureau to participate in the American Housing Survey (AHS). The Census Bureau conducts the AHS to help federal, state, and local governments, businesses, and non-profits understand changes in housing conditions, costs, and other aspects of your community to better design policies and services like schools, roads, and hospitals. Soon, a Census Bureau employee from your area will contact you to collect your response to the AHS.

Here's what you should know:

- **Every two years, a select number of households are chosen to represent thousands of others like theirs**. Don't miss this opportunity to have your voice heard!
- **The Census field representative will show you their badge when they visit your home**. If they call you, they will provide their name and Regional Office information to confirm employment with the Census Bureau.
- **U.S. householders paid a median of $1,037 per month on housing costs in 2017**. You can look at the enclosed data wheel to see how your area compares.
- **Answers to the most frequently asked questions about this survey are included on the back of this letter**. If you have other questions or want to learn more about the survey, please go to <www.census.gov/programs-surveys/ahs/about/respondent-information.html> or call your U.S. Census Bureau Regional Office at 1–800–992–3530.

Our commitment to you:

We promise not to publicly release your responses in a way that could identify you.

We promise that we will use every technology, statistical methodology, and physical security procedure available to protect your information.

Thank you in advance for your help with this important survey.

Sincerely,

Steven D. Dillingham
Director

La Oficina del Censo de los EE. UU. se encuentra realizando la Encuesta de Viviendas de los Estados Unidos (AHS) en su comunidad y su direcci n fue escogida para participar. Sus respuestas son confidenciales y, pronto, un representante local se pondr en contacto con usted para completar la encuesta. Si usted prefiere que la entrevista se realice en espa ol, comun quese con nosotros al 1–800–992–3530 o por correo electr nico a Los.Angeles.Regional.Office@census.gov o escr banos a esta direcci n: Regional Director, U.S. Census Bureau, 15350 Sherman Way Ste 400, Van Nuys, CA 91406-4203.

United States Census Bureau

census.gov

WHAT IS THIS SURVEY ABOUT?

The American Housing Survey (AHS) collects up-to-date information on housing quality and costs in the United States. To measure housing quality, the survey includes questions about equipment breakdowns, leaks, and other problems. To measure housing costs, the survey includes questions about mortgage and rental costs, utility costs, and repair and remodeling costs. The AHS also asks about your household, including demographic and income questions. Combining household information with housing quality and cost information helps to measure the housing challenges faced by homeowners and renters. The information also helps to measure important changes in our housing stock as it ages, and when it is eventually remodeled or replaced.

HOW WAS I SELECTED FOR THIS SURVEY?

The U.S. Census Bureau chose your address, not you personally, to participate in this survey. We randomly selected a sample of addresses throughout the United States. We need a response from every home in our sample to get a complete picture of housing quality and costs across the country. Your answers represent not only your home, but also thousands of other homes like yours. If you move, this address will stay in the survey, and we will interview the household that moves here.

I COMPLETED THIS SURVEY BEFORE. WHY ARE YOU ASKING ME TO DO IT AGAIN?

It is important that someone at this address completes the survey again so that we may measure the changes (or lack of changes) in the number of homes available in the United States, the physical condition of the housing, and the characteristics of the occupants. Some addresses in the AHS are asked to complete the survey once every two years. Other addresses are asked to complete the survey every few years.

I THOUGHT THAT THE CENSUS BUREAU OPERATED ONLY EVERY TEN YEARS WHEN IT COUNTED PEOPLE. WHAT IS THE CENSUS BUREAU DOING NOW?

Besides the decennial census conducted every ten years, we collect many different kinds of statistics through other censuses and surveys. We conduct other censuses at regular intervals, including the Economic Census and the Census of Governments. In addition, we conduct various surveys to collect data on a monthly basis in order to provide current information on unemployment rates, retail and wholesale trade, various manufacturing activities, new housing construction, and a number of other topics. Also, we conduct annual surveys on business, manufacturing, governments, family income, health, and education. You may also encounter the Census Bureau conducting collections on behalf of other agencies, like the AHS, which the Census Bureau conducts for the U.S. Department of Housing and Urban Development (HUD).

IS THIS SURVEY AUTHORIZED BY LAW? WHAT PROTECTION DO I HAVE?

The U.S. Department of Housing and Urban Development is authorized to collect this information under the Housing and Urban-Rural Recovery Act of 1983 (12 U.S.C. 1701z-1, 1701z-2(g), and 1701z-10a). The Census Bureau conducts the survey on behalf of HUD under the authority of 13 U.S.C. 8(b). The Census Bureau is required by law to protect your information. The Census Bureau is not permitted to publicly release your responses in a way that could identify this household. Federal law protects your privacy and keeps your answers confidential (Title 13, United States Code, Section 9(a)). Your answers may be combined with information that you give to other agencies. By law, the Census Bureau can only use your responses for statistical research. For more information, please visit the Census Bureau's Web site on combining data: <www.census.gov/about/what/admin-data.html>. Per the Federal Cybersecurity Enhancement Act of 2015, your data are protected from cybersecurity risks through screening of the systems that transmit your data. Disclosure of the information provided to us is permitted under the Privacy Act of 1974 (5 U.S.C. § 552a) and may be shared with other Census Bureau staff for the work-related purposes identified in this statement. Disclosure of this information is also subject to the published routine uses as identified in the Privacy Act System of Records Notice COMMERCE/Census-3, Demographic Survey Collection (Census Bureau Sampling Frame). Furnishing this information is voluntary. Failure to provide this information may affect the Census Bureau's ability to collect information on U.S. housing quality and costs.

HOW LONG WILL IT TAKE?

We estimate that completing the AHS will take 40 minutes on average. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to the Director, Housing and Demographic Analysis Division, Office of Policy Development and Research, Office of Economic Affairs, Department of Housing and Urban Development, Washington, DC 20410. This information collection is authorized by OMB control 2528-0017 (expires May 31, 2020). If this number were not displayed, we could not conduct this survey.

AHS-26/66(L) (4-2019)

Treatment: 2, 5, or 10 Dollar Bill

AHS-26/66(L) (Los Angeles)

UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001
OFFICE OF THE DIRECTOR

Thank you in advance for your participation in this survey.
Please accept this token of our appreciation.

Dear Resident,

Your address has been selected by the U.S. Census Bureau to participate in the American Housing Survey (AHS). The Census Bureau conducts the AHS to help federal, state, and local governments, businesses, and non-profits understand changes in housing conditions, costs, and other aspects of your community to better design policies and services like schools, roads, and hospitals. Soon, a Census Bureau employee from your area will contact you to collect your response to the AHS.

Here's what you should know:

- **Every two years, a select number of households are chosen to represent thousands of others like theirs**. Don't miss this opportunity to have your voice heard!
- **The Census field representative will show you their badge when they visit your home**. If they call you, they will provide their name and Regional Office information to confirm employment with the Census Bureau.
- **U.S. householders paid a median of $1,037 per month on housing costs in 2017**. You can look at the enclosed data wheel to see how your area compares.
- **Answers to the most frequently asked questions about this survey are included on the back of this letter**. If you have other questions or want to learn more about the survey, please go to <www.census.gov/programs-surveys/ahs/about/respondent-information.html> or call your U.S. Census Bureau Regional Office at 1–800–992–3530.

Our commitment to you:

We promise not to publicly release your responses in a way that could identify you.

We promise that we will use every technology, statistical methodology, and physical security procedure available to protect your information.

Thank you in advance for your help with this important survey.

Sincerely,

Steven D. Dillingham
Director

La Oficina del Censo de los EE. UU. se encuentra realizando la Encuesta de Viviendas de los Estados Unidos (AHS) en su comunidad y su direcciï n fue escogida para participar. Sus respuestas son confidenciales y, pronto, un representante local se pondrÛ en contacto con usted para completar la encuesta. Si usted prefiere que la entrevista se realice en espaæol, comunÞquese con nosotros al 1–800–992–3530 o por correo electrï nico a Los.Angeles.Regional.Office@census.gov o escrÞbanos a esta direcciï n: Regional Director, U.S. Census Bureau, 15350 Sherman Way Ste 400, Van Nuys, CA 91406-4203.

United States Census Bureau

census.gov

WHAT IS THIS SURVEY ABOUT?

The American Housing Survey (AHS) collects up-to-date information on housing quality and costs in the United States. To measure housing quality, the survey includes questions about equipment breakdowns, leaks, and other problems. To measure housing costs, the survey includes questions about mortgage and rental costs, utility costs, and repair and remodeling costs. The AHS also asks about your household, including demographic and income questions. Combining household information with housing quality and cost information helps to measure the housing challenges faced by homeowners and renters. The information also helps to measure important changes in our housing stock as it ages, and when it is eventually remodeled or replaced.

HOW WAS I SELECTED FOR THIS SURVEY?

The U.S. Census Bureau chose your address, not you personally, to participate in this survey. We randomly selected a sample of addresses throughout the United States. We need a response from every home in our sample to get a complete picture of housing quality and costs across the country. Your answers represent not only your home, but also thousands of other homes like yours. If you move, this address will stay in the survey, and we will interview the household that moves here.

I COMPLETED THIS SURVEY BEFORE. WHY ARE YOU ASKING ME TO DO IT AGAIN?

It is important that someone at this address completes the survey again so that we may measure the changes (or lack of changes) in the number of homes available in the United States, the physical condition of the housing, and the characteristics of the occupants. Some addresses in the AHS are asked to complete the survey once every two years. Other addresses are asked to complete the survey every few years.

I THOUGHT THAT THE CENSUS BUREAU OPERATED ONLY EVERY TEN YEARS WHEN IT COUNTED PEOPLE. WHAT IS THE CENSUS BUREAU DOING NOW?

Besides the decennial census conducted every ten years, we collect many different kinds of statistics through other censuses and surveys. We conduct other censuses at regular intervals, including the Economic Census and the Census of Governments. In addition, we conduct various surveys to collect data on a monthly basis in order to provide current information on unemployment rates, retail and wholesale trade, various manufacturing activities, new housing construction, and a number of other topics. Also, we conduct annual surveys on business, manufacturing, governments, family income, health, and education. You may also encounter the Census Bureau conducting collections on behalf of other agencies, like the AHS, which the Census Bureau conducts for the U.S. Department of Housing and Urban Development (HUD).

IS THIS SURVEY AUTHORIZED BY LAW? WHAT PROTECTION DO I HAVE?

The U.S. Department of Housing and Urban Development is authorized to collect this information under the Housing and Urban-Rural Recovery Act of 1983 (12 U.S.C. 1701z-1, 1701z-2(g), and 1701z-10a). The Census Bureau conducts the survey on behalf of HUD under the authority of 13 U.S.C. 8(b). The Census Bureau is required by law to protect your information. The Census Bureau is not permitted to publicly release your responses in a way that could identify this household. Federal law protects your privacy and keeps your answers confidential (Title 13, United States Code, Section 9(a)). Your answers may be combined with information that you give to other agencies. By law, the Census Bureau can only use your responses for statistical research. For more information, please visit the Census Bureau's Web site on combining data: <www.census.gov/about/what/admin-data.html>. Per the Federal Cybersecurity Enhancement Act of 2015, your data are protected from cybersecurity risks through screening of the systems that transmit your data. Disclosure of the information provided to us is permitted under the Privacy Act of 1974 (5 U.S.C. § 552a) and may be shared with other Census Bureau staff for the work-related purposes identified in this statement. Disclosure of this information is also subject to the published routine uses as identified in the Privacy Act System of Records Notice COMMERCE/Census-3, Demographic Survey Collection (Census Bureau Sampling Frame). Furnishing this information is voluntary. Failure to provide this information may affect the Census Bureau's ability to collect information on U.S. housing quality and costs.

HOW LONG WILL IT TAKE?

We estimate that completing the AHS will take 40 minutes on average. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to the Director, Housing and Demographic Analysis Division, Office of Policy Development and Research, Office of Economic Affairs, Department of Housing and Urban Development, Washington, DC 20410. This information collection is authorized by OMB control 2528-0017 (expires May 31, 2020). If this number were not displayed, we could not conduct this survey.

AHS-26/66(L) (4-2019)

U.S. Department of Housing and Urban Development
Office of Policy Development and Research
Washington, DC 20410-6000

December 2023