DRAFT: 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

Last Updated: September 25, 2020

Contents

Executive Summary

PLEASE NOTE: Estimates in this draft memorandum are subject to change. In particular, as we note below, the variance estimates used for statistical inference in analyses of nonresponders should be treated with care, as replicate weights for nonresponders are unavailable.

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.[1] 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.[2]

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.[3] 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,

[1] See the OMB guidance at: https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/assets/OMB/inforeg/statpolicy/standards_stat_surveys.pdf.

[2] 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.

[3] In Section 1.1, we discuss distinctions between different types of units in each categories—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

Table 1: Non-response 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)

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.

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 nationally-representative 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.[4]

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 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

[4] 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).

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.[5]

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:

1. **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.

2. **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.

[5] 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.

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.[6]

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 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 non-vacant interviews[7] that are not part of the bridge or metropolitan samples.

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

[7] 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 tem-

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.[8] 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 (gray, `BASEWGT`) and bias-adjusted (green, `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. 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).

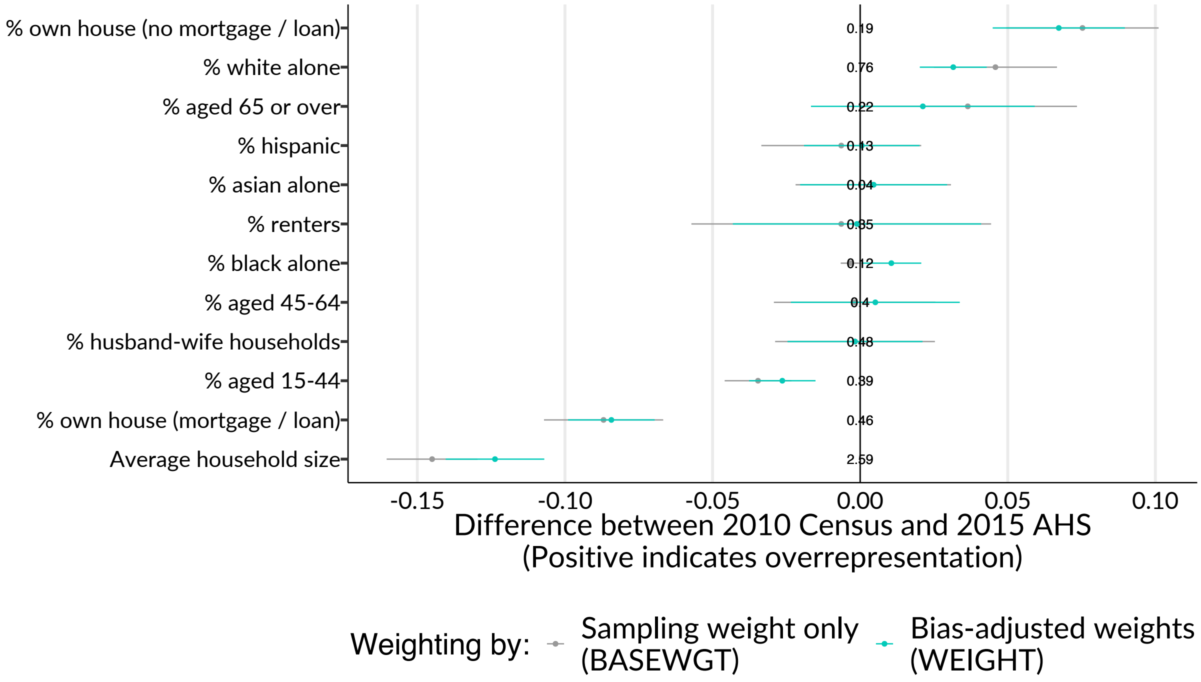
The number centered at zero on the first row indicates that 19 percent of householders interviewed in the 2010 Census owned their house outright (without loan or mortgage). Taking the analysis at face value, the gray point on this row indicates that the 2015 AHS contains roughly eight percentage points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

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.

porarily 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.

[8] 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.

Figure 1: Divergence between the 2010 Census and national estimates derived from the 2015 AHS. Points correspond to the difference between 2010 Census figures and sample-weighted (gray) and bias-adjusted (green) 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-wifehouseholds) 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 15-44 and large households are underrepresented in the AHS sample.

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, 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.[9] 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:

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

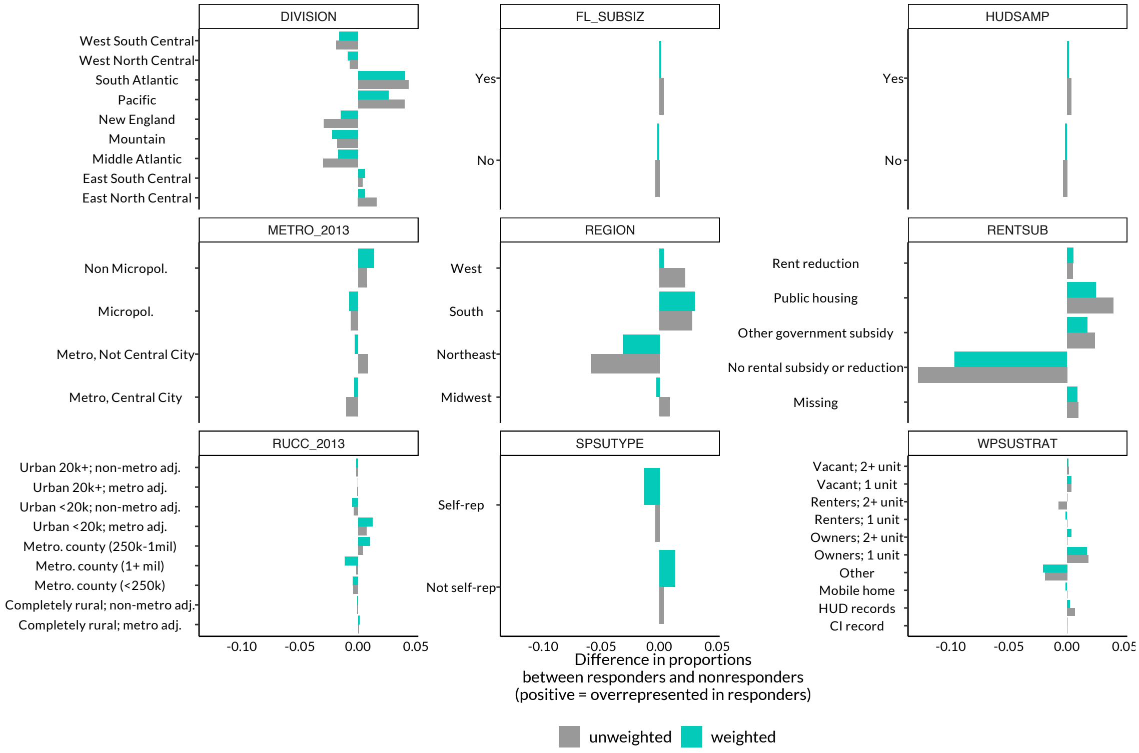
Results

Figure 2, which focuses on the 2019 wave, shows how responders differ from nonresponders for several sampling frame attributes. The gray bars represent the point estimates without weighting; the green 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. Finally, even after this initial reweighting, the two groups still look significantly different.[10] Results for the 2015 and 2017 waves look substantively similar, and all differences were statistically significant at the $p < 0.001$ level.

[9] Note that this test does not take account of the clustering and stratification involved in the sampling design and makes an anti-conservative assumption of independent sampling.

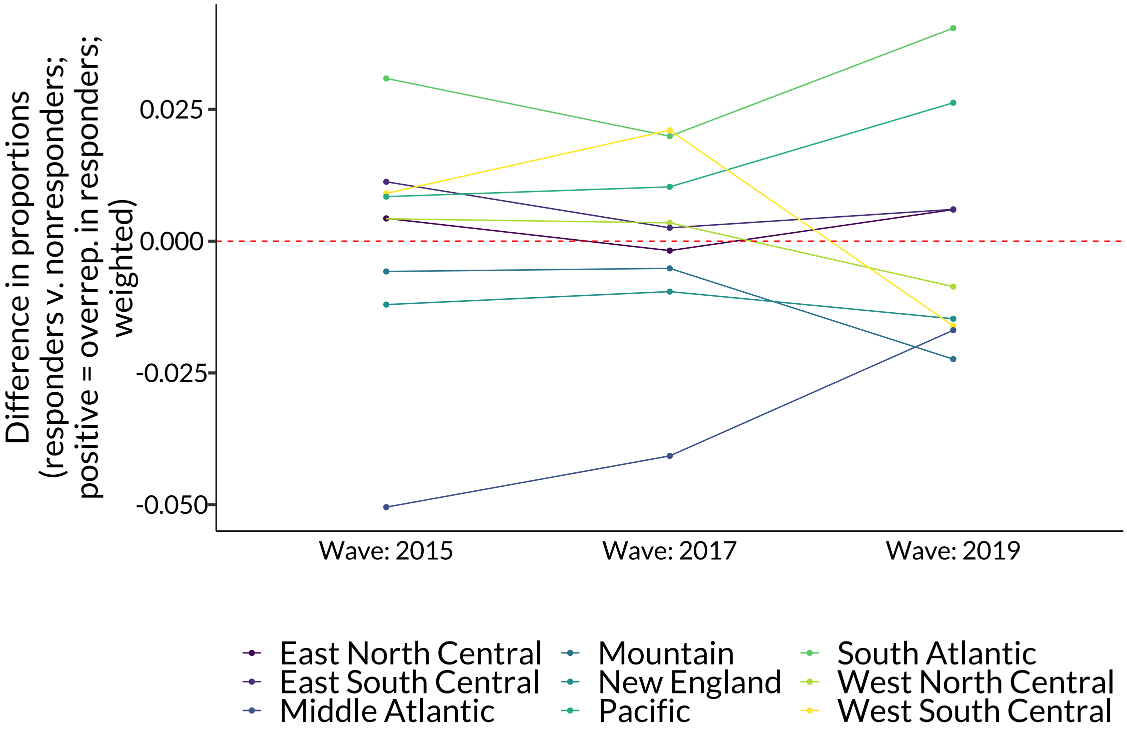
[10] Appendix Table 7 shows the p-values for each of the tests.

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.

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, re-parametrize 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 re-estimate 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.[11] These ACS

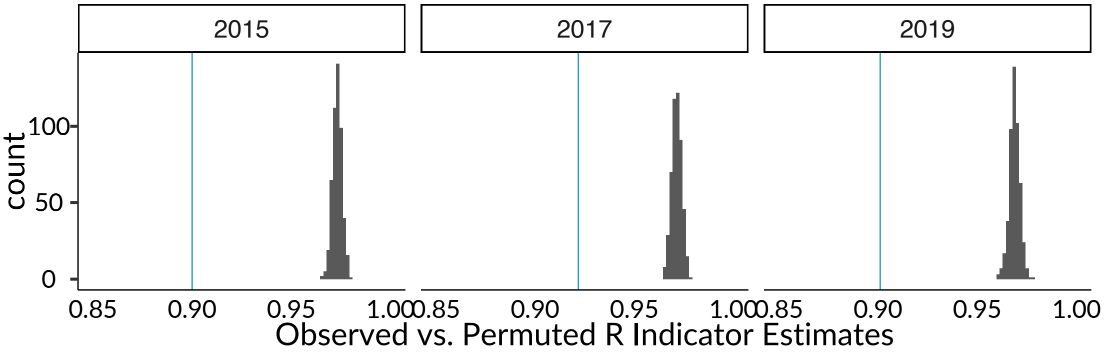
[11] Some predictors had to be dropped due to collinearity, which arises when two or more variables contain very similar

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 (dark gray 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 8 in Appendix Section 2.3 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 8 in the appendix, confirm the main finding: statistically significant evidence that nonresponders' attributes differ from those of responders.

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. 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

or equivalent information on units in the analysis, and thereby "cancel out" each other's estimated influence on the outcome.

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:

1. **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.
2. **Refusal:**[12]
   - `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.[13] 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 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

[12] 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.

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

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.[14]

[14] 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.[15]

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

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

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

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 4 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).[16] 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.

[16] The choice of threshold can be calibrated to balance precision with recall. The results presented use sklearn's auto 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).

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.
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.[17] 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.

[17] 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.

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

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.

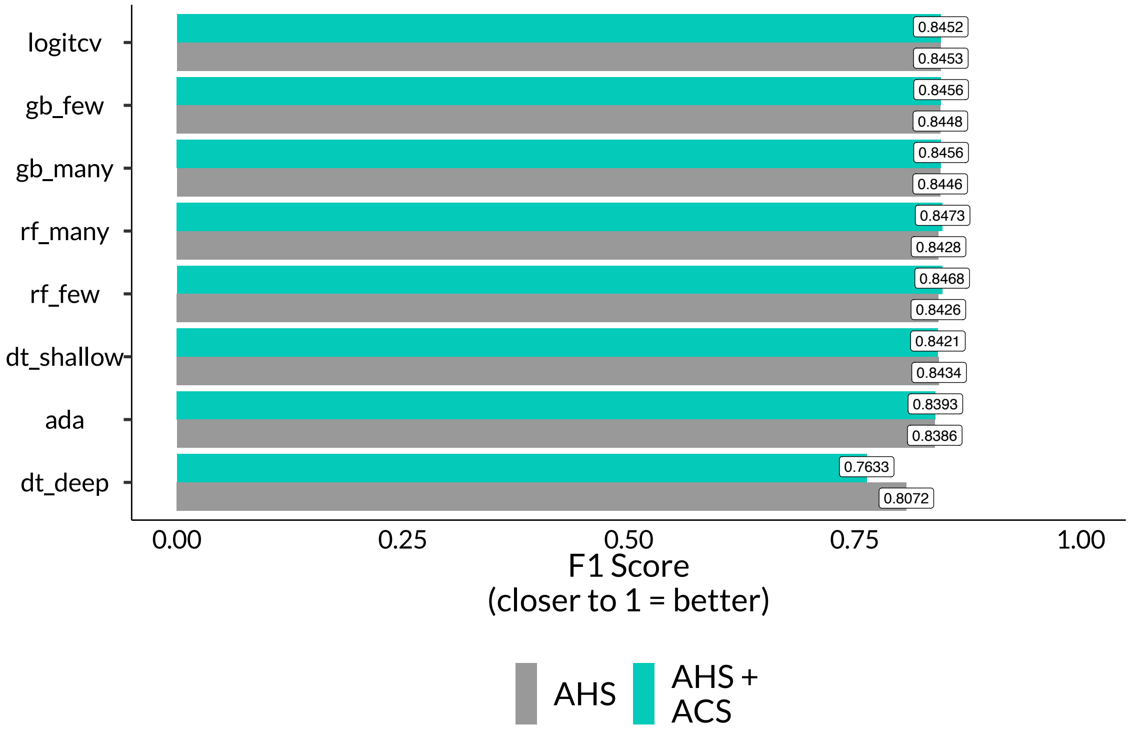
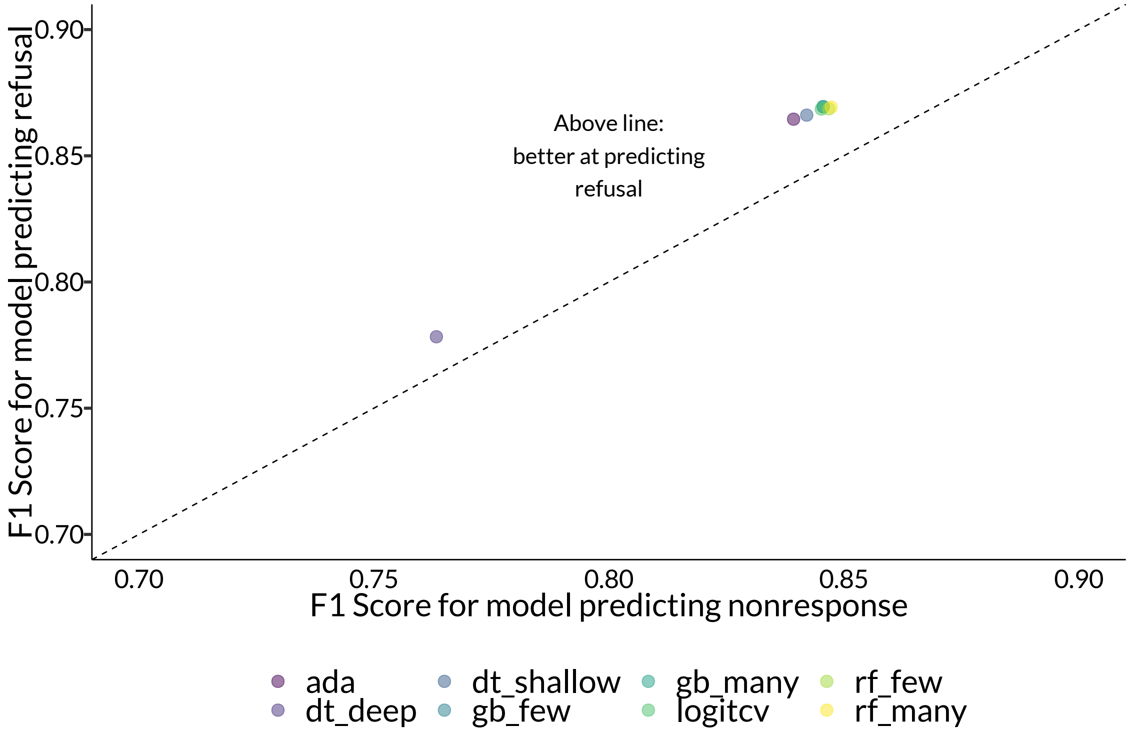


Figure 6 focuses on the contrast between our ability to predict nonresponse (previous graph) and our ability to predict refusal. Each dot represents one of the final models. The fact that all models are above the 45 degree line shows that while the predictions of nonresponse and refusal each have high 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,[18] we might be better able to leverage past refusal to predict refusal in a focal wave than past nonresponse to predict nonresponse in a focal wave.

[18] As we discuss in Section 1.1, 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.

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 non-refusal) sign.

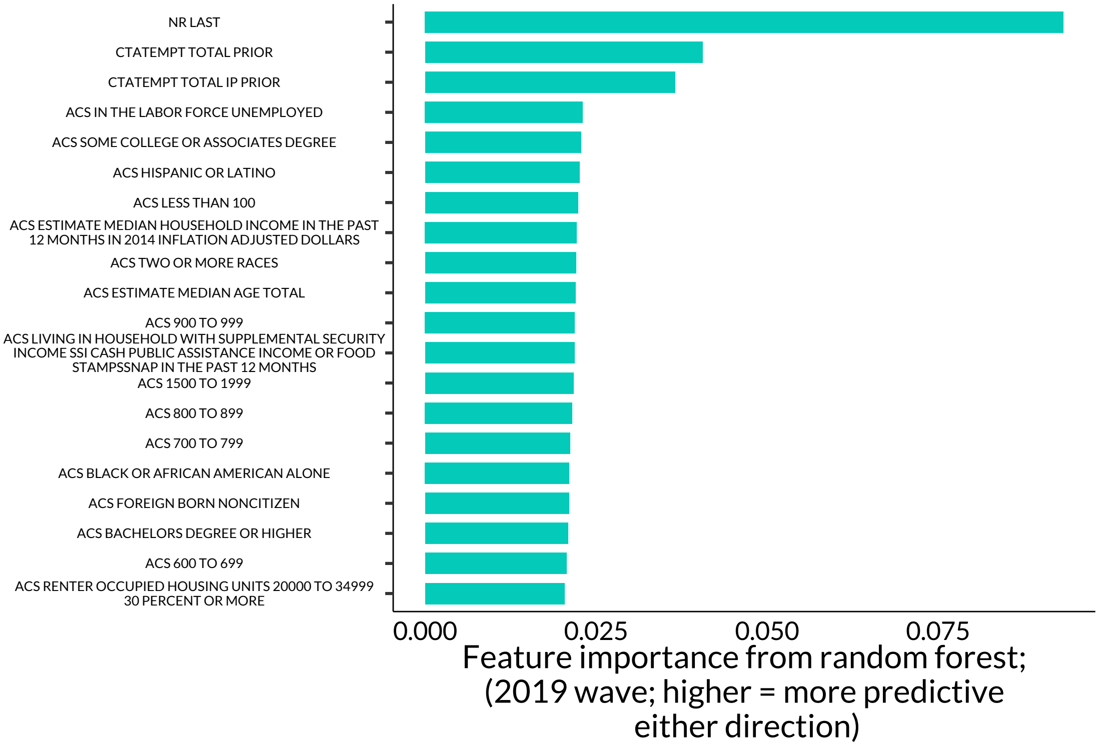
Importantly, and as in the panel attrition analysis we discuss later, all features are *predictive* rather than *causal*. For instance, there might be unobserved characteristics of a unit that lead that unit to refuse in both the 2015 and 2017 waves. The "did not respond in 2015" feature in the model 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.[19] 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.[20]

Figure 8 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 24 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.

[19] 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.

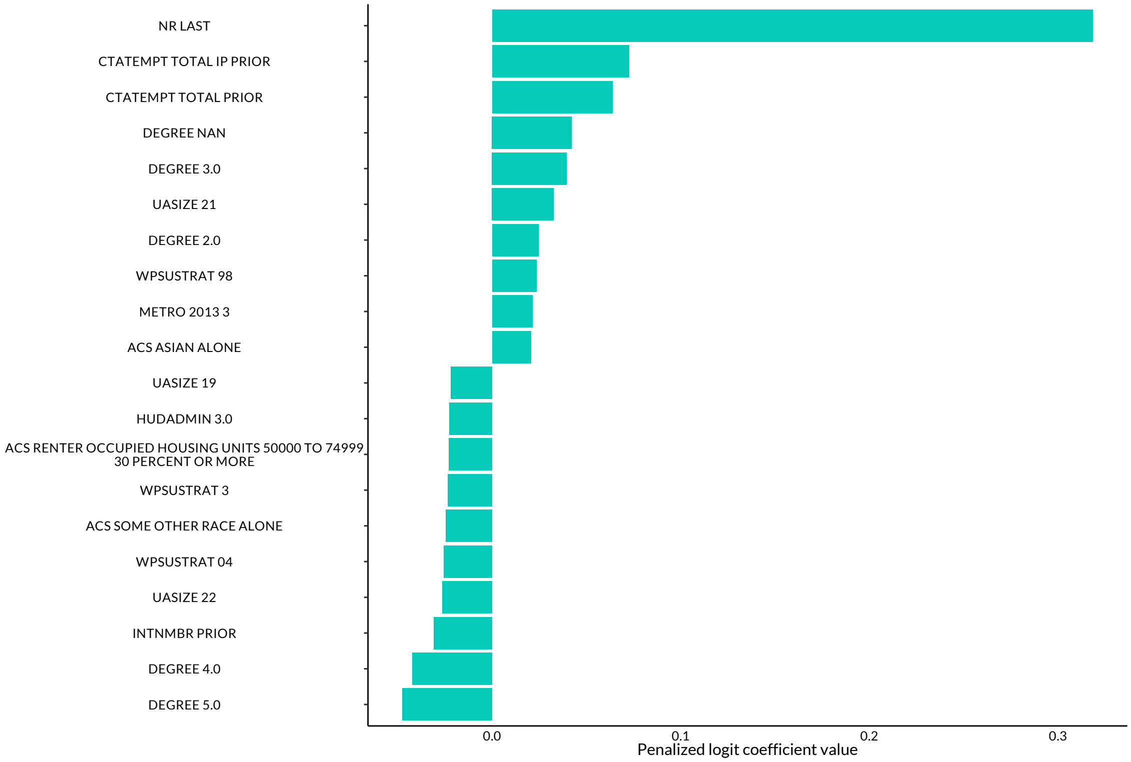
[20] The pattern—area-level lower SES is associated with a higher likelihood of unit-level nonresponse (e.g., Maitland et al. 2017)—has been observed in other social surveys. While trust is one mechanism, there might be many others like work schedules, time pressures, and more.

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 non-directional 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.

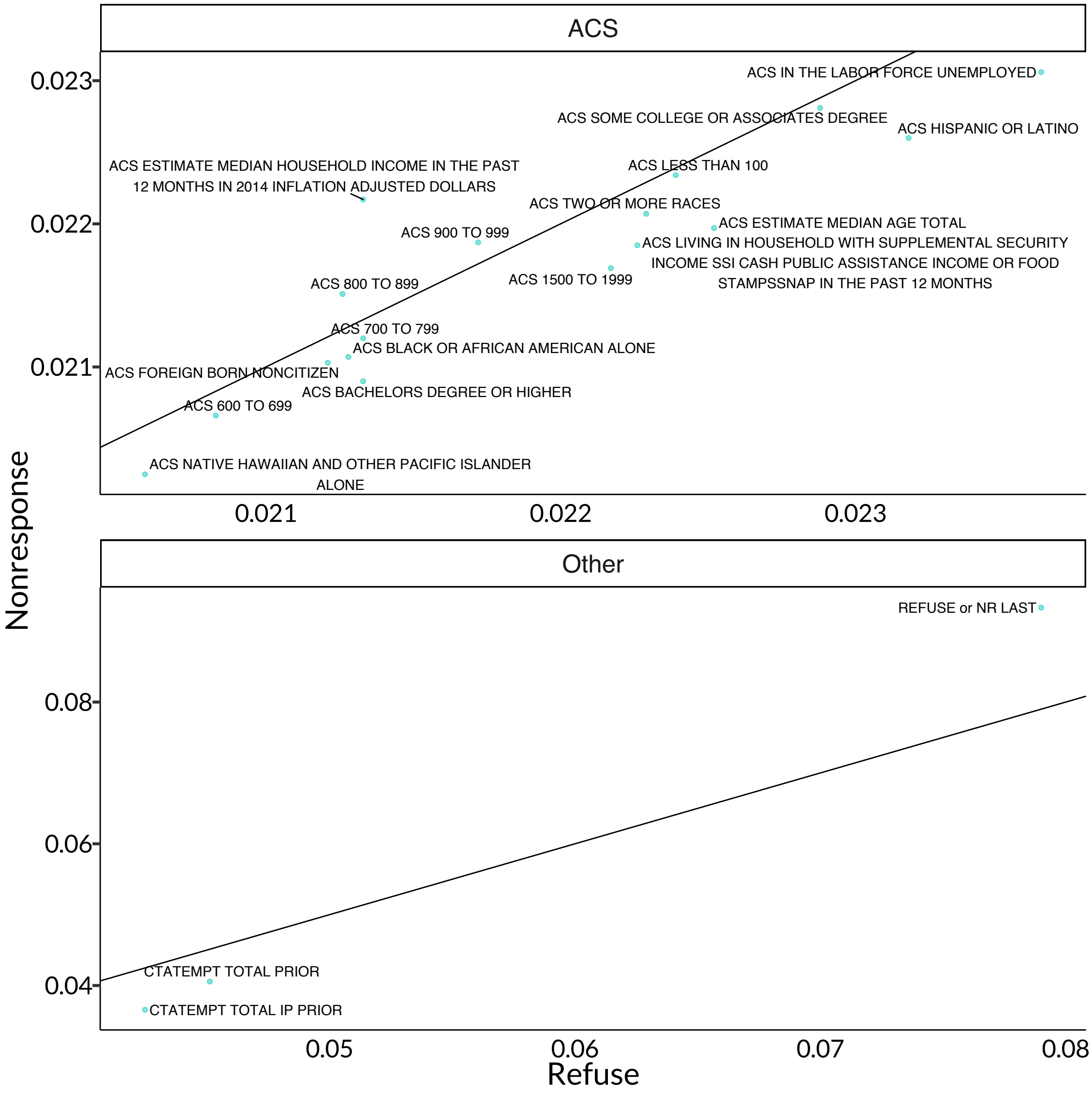
Figure 8: Most important predictors of nonresponse: penalized logit; 2019 wave The figure illustrates the top 10 positive and top 10 negative features.



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 non-behavioral 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,[21] 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 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.[22] Figure 11 shows the top 20 items, among those not imputed, that have the highest rate of nonreport. Focusing on items with high rates

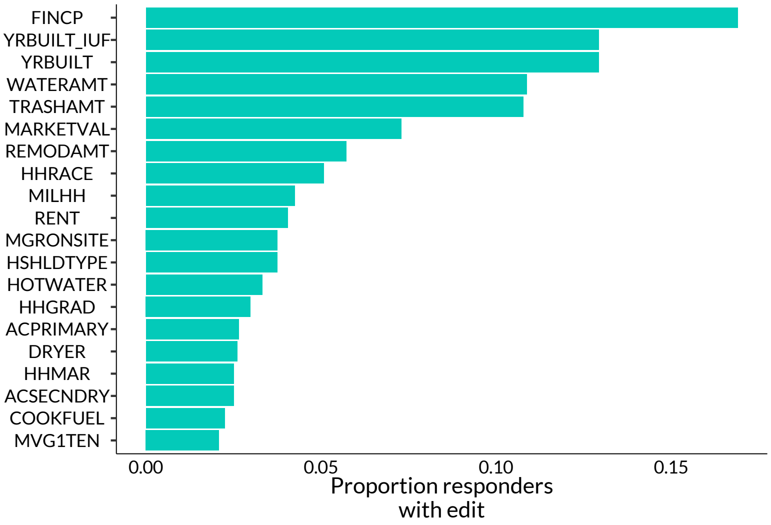
[21] In particular, sklearn classifiers vary in whether they accept a sample_weights argument, making it more straightforward to first estimate a range of classifiers and then choose the top-performing one that also accepts survey weights.

[22] 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.

of nonreport, we see some patterns like potentially-sensitive items about neighborhood safety or financial challenges.[23] For instance, the following high missingness items might be more sensitive:

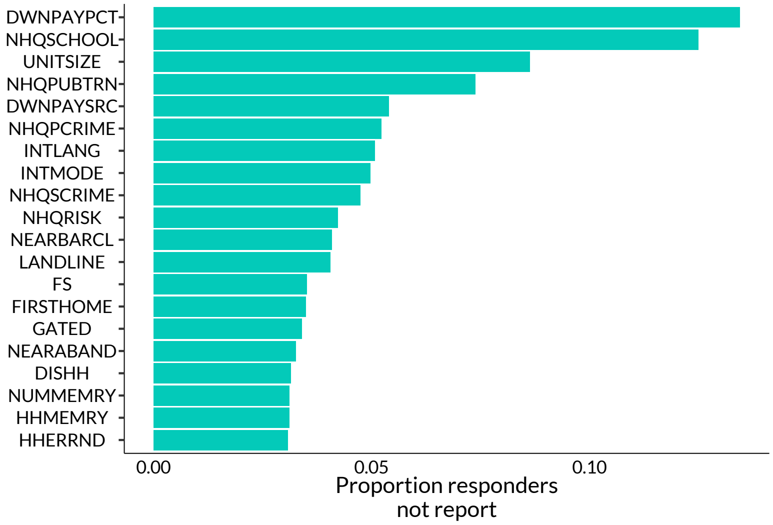
1. `NHQPCRIME`: Agree or Disagree: This neighborhood has a lot of petty crime
2. `NHQSCRIME`: Agree or Disagree: This neighborhood has a lot of serious crime
3. `NUMMEMRY`: Number of persons living in this unit who have difficulty concentrating or remembering

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).



[23] We also see others like interview mode and language that might be not reported for more survey administration-based 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 missigness could stem from two sources:

- Missingness due to the item itself—for instance, sensitive questions having higher missingness; or
- 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:

1. The unit's identifier, and
2. 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.

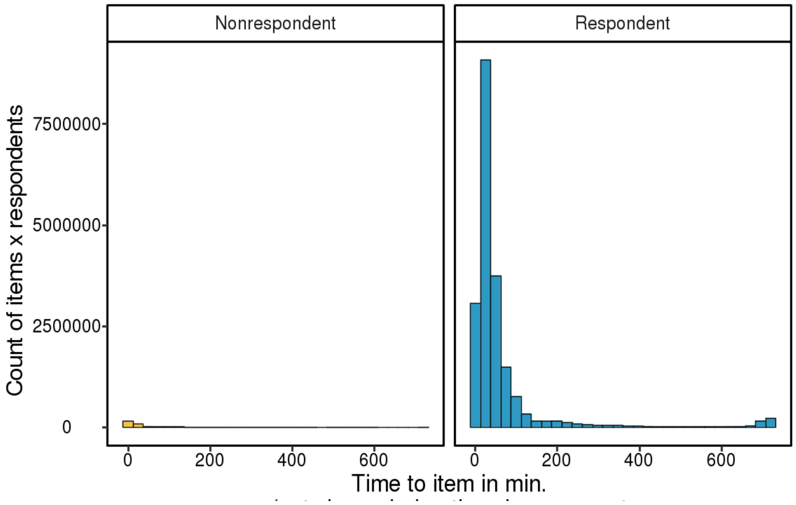
The parsing process was not perfect. In particular, for 18 units (less than 0.001 percent of the total occupied interviews included in the analysis), there were issues in how the timestamps were recorded that led us to exclude them from the analysis.

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:

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

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.

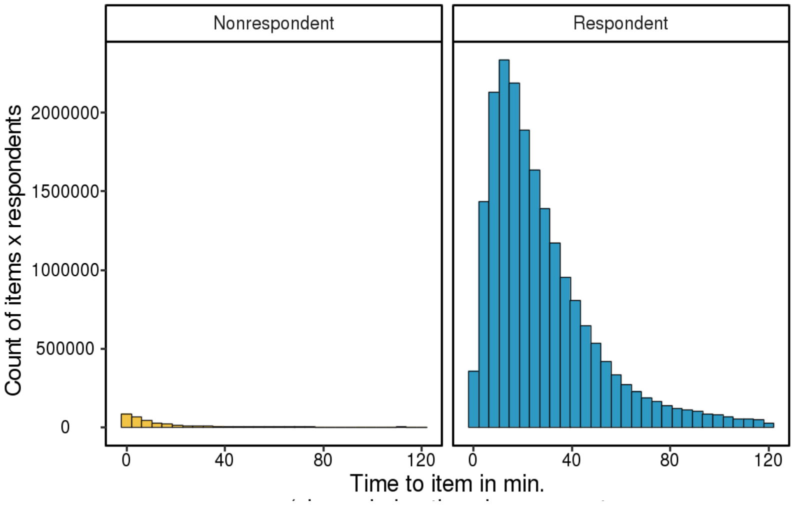
Figure 12: Distribution of item-specific relative durations (raw) The second small peak at closer to 8-10 hours shows that some respondents had multiple distinct sessions in the same day.



Due to this challenge, we use a rough measure of the start of the survey—the keystroke indicating the initiation of a new survey.[24] 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.

[24] The action of "Enter Field" on `STARTCP`.

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:

$$\text{Do not respond to item (1 = yes)}_{ik} = \alpha + \beta_1 \text{Relative duration}_{ik} + \gamma_i + \delta_k + \epsilon_{ik}.$$

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:

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

2. *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.

3. *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.

4. *We then merge the information on the respondent's survey item response status to the relative duration of that item for that particular respondent*[25]

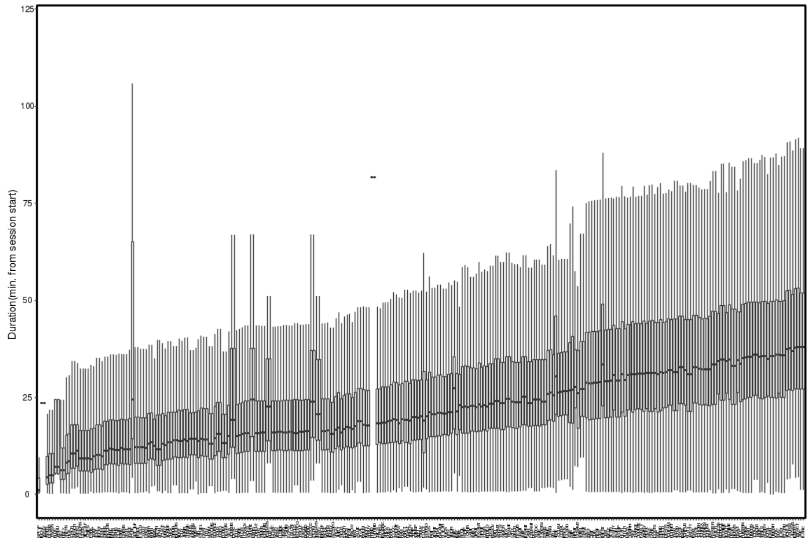
5. *Since we do not observe all items in the trace file for each responder, we create two versions of the duration variable*: one with the values from above; the other that imputes respondents missing duration for a particular item to the mean duration for that item. We filter out item-respondent pairs 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.[26] 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

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.



[25] 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.

[26] Since our outcome variable is nonresponse to the item.

Overall, the parsing shows how some items that might be more sensitive like the perceptions of neighborhood items are also towards 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 duration	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.[27] 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.[28] 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 non-missing responses, and include in the set of potential predictors a dummy variable for each numeric variable, indicating whether mean imputation was employed. 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.

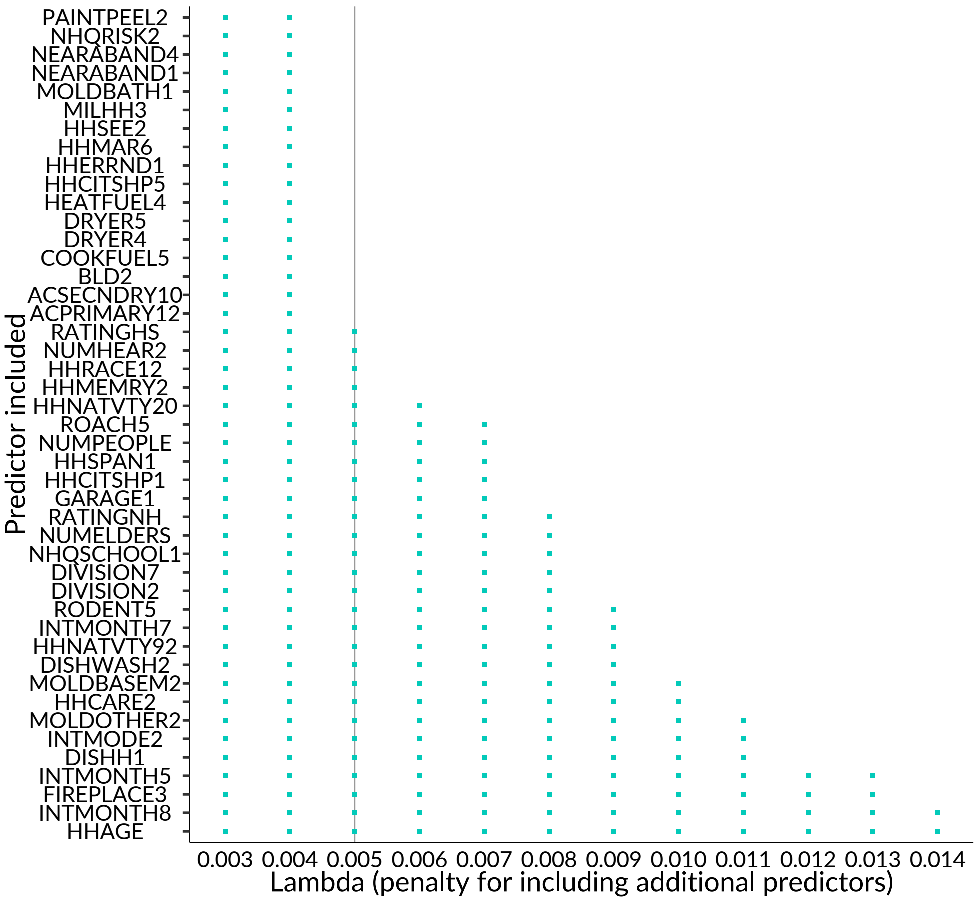
Results

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.

[27] 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.

[28] 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.

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 too 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 squares 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.

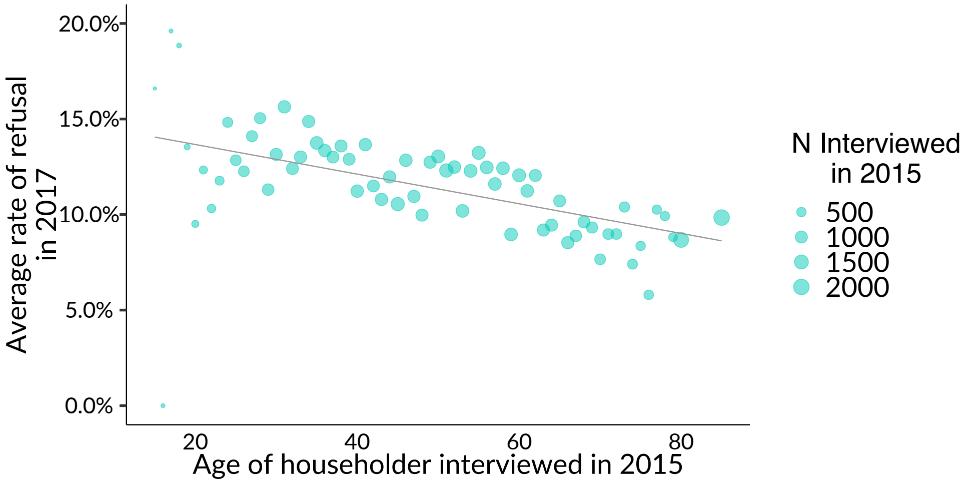
Table 6 presents the 15 predictors that the lasso is least likely to drop as the penalty increases (e.g., those at the bottom of Figure 15).

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 Census Bureau, and employ Fay's BRR method (see footnote 8 above).

term	estimate	std.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

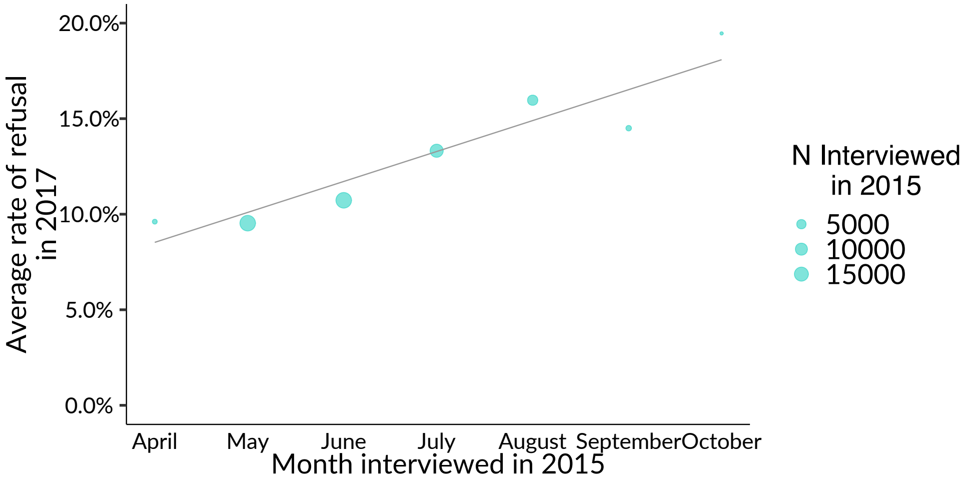
Turning to the first coefficient, `HHAGE`, we see that age is negatively correlated with the probability of refusal. As depicted on 16, the relationship is approximately linear: as the age of the householder interviewed in 2015 decreases, so too does the probability of that household refusing to do the 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.[29]

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 non-refusers.[30] 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.

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.

[29] 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 25 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.

[30] 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.

5.1 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 multi-wave 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.[31] We examine the relationship between the household total income (HINCP) and whether the respondent lives in inadequate housing.[32] 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 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.[33]

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. Blue lines correspond to units who responded both in 2015 and

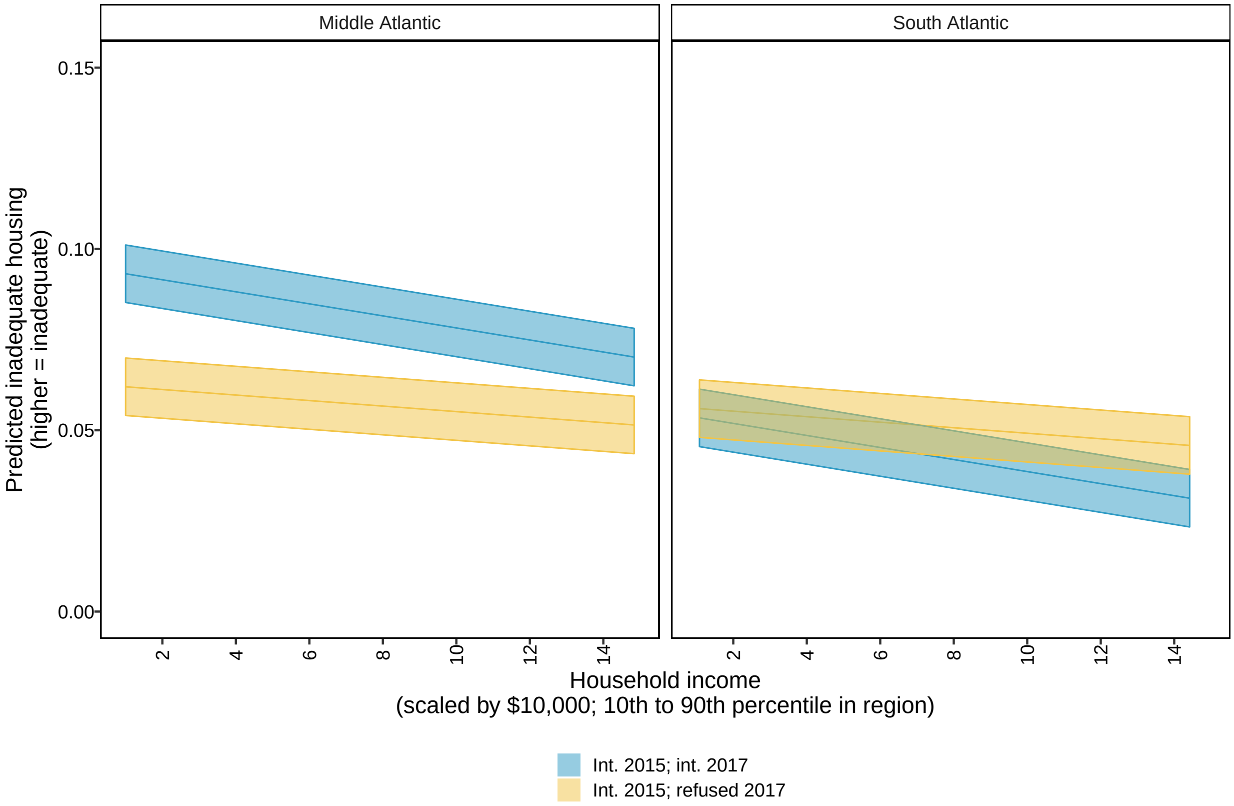
[31] These are based on the NOINT variable in the case history file.

[32] 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 non-scaled version.

[33] 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 ($p = 0.03$).

in 2017 while yellow 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 non-attritors with similar income levels. Researchers who restrict an analysis of longitudinal trends to households that appear in both waves would essentially only estimate the blue line, ignoring the yellow. 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.

Figure 18: Predicted inadequacy by income in 2015: respondents who then refuse in 2017 versus respondents who respond both waves The analysis constucts 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 metro-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.

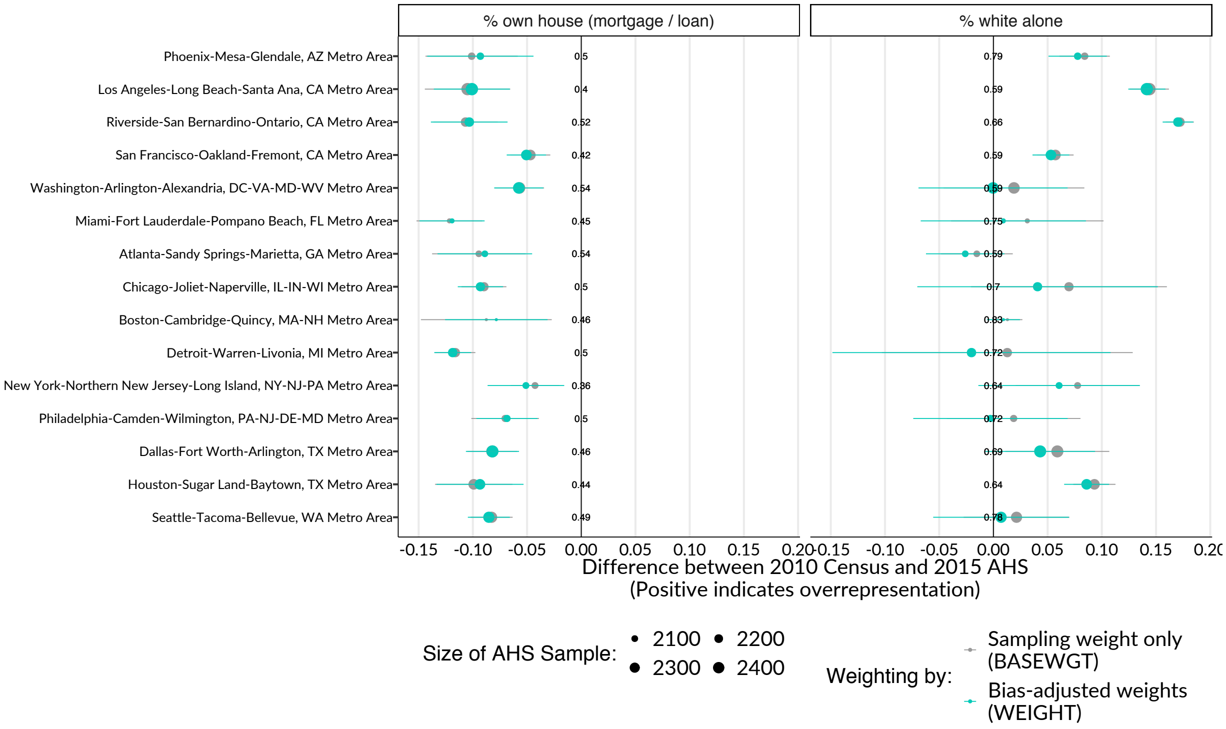
Methods

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: Metro-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. “Replication 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 Information 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, Weighting, and Error Estimation.* Technical report. https://www2.census.gov/programs-surveys/ahs/2015/.

A Appendix

A.1 Additional results from the chi-squared analysis

Table 7: P-values from chi-squared analysis of differences between responders and nonresponders All differences are significant at the $p < 0.001$ level.

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; non-metro adj.	-0.0009	p $<$ 0.001
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; non-metro adj.	-0.0050	p $<$ 0.001
RUCC_2013	Urban 20k+; metro adj.	-0.0001	p $<$ 0.001
RUCC_2013	Urban 20k+; non-metro 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

Figure 20: Differences between responders and nonresponders: 2017 wave.

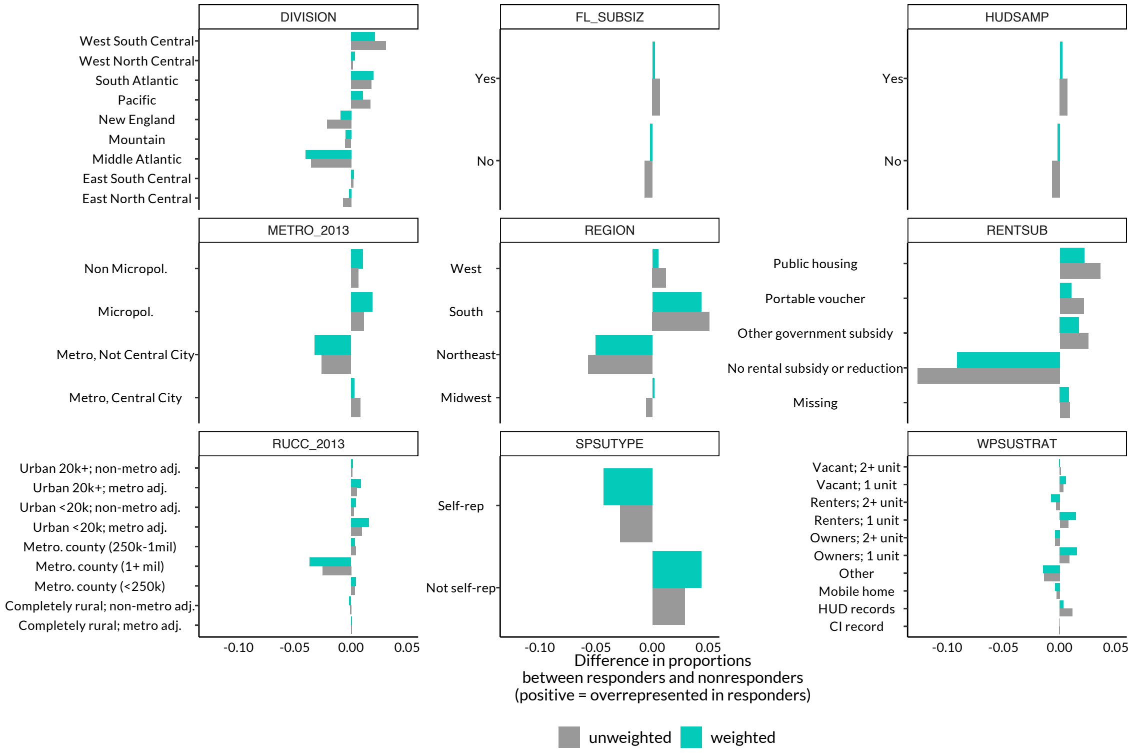
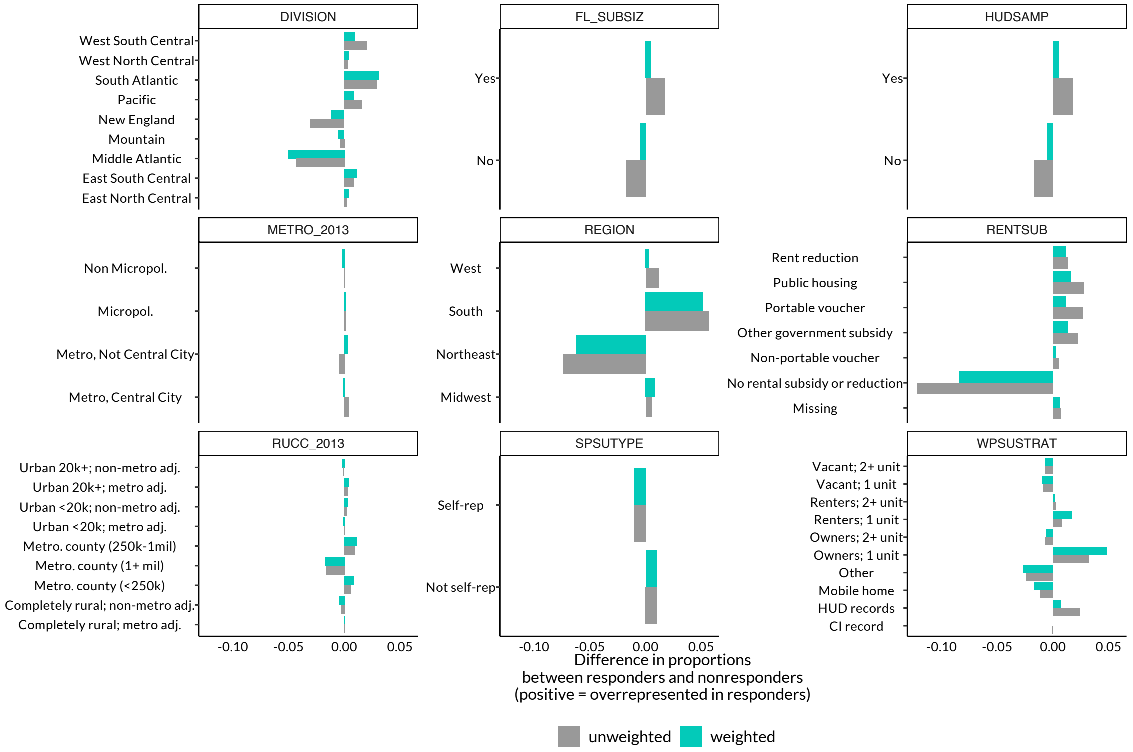


Figure 21: Differences between responders and nonresponders: 2015 wave.



A.2 Additional results from R indicator analysis

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

Table 8: Results from R-indicator analysis.

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
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.

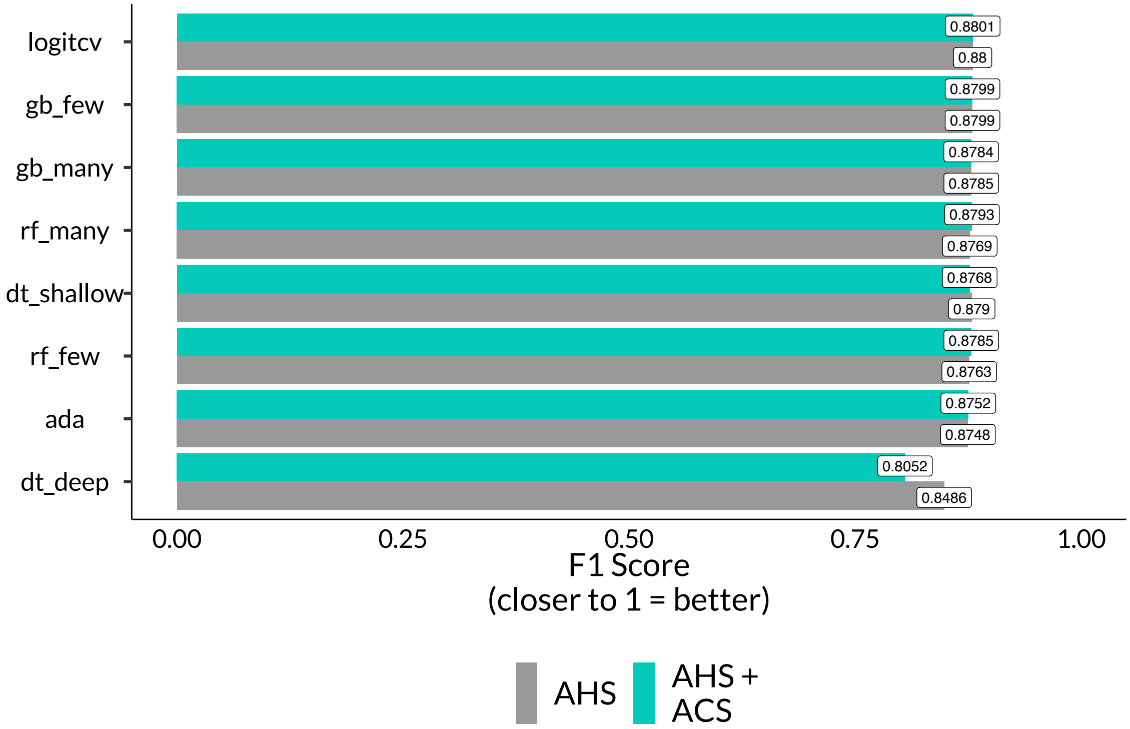


Figure 23: Ability to predict refusal: 2017 and 2019 wave.

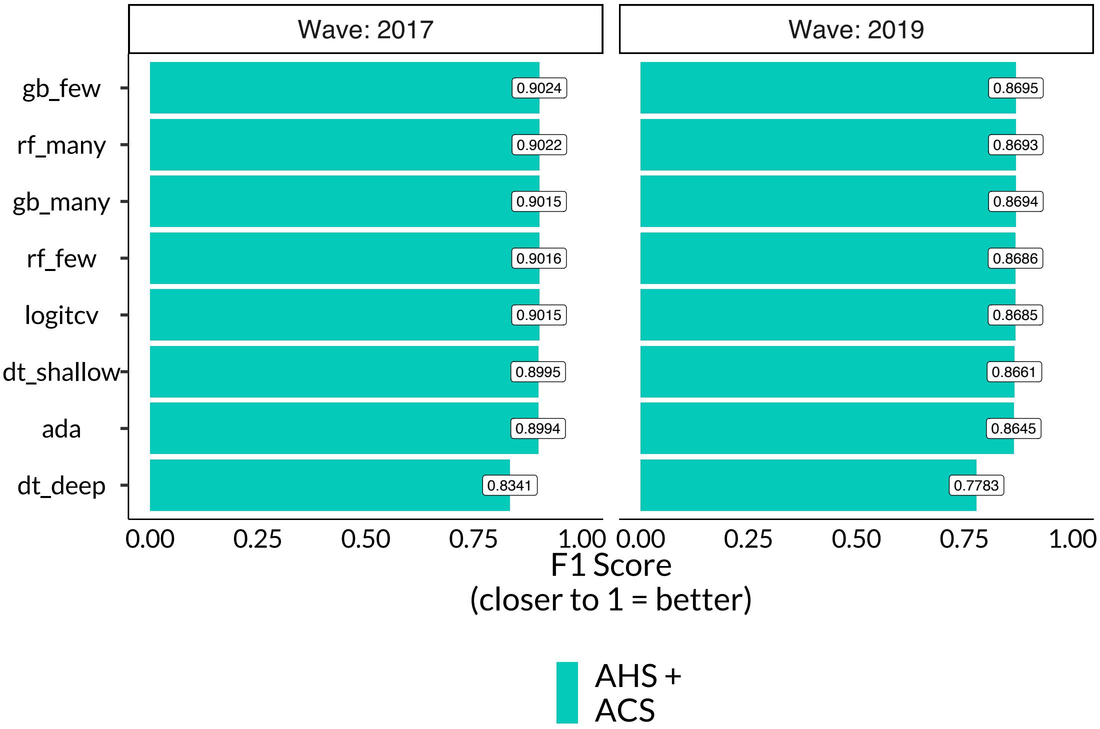


Figure 24: Remaining feature importances outside of the top 20: random forest; 2019 wave.

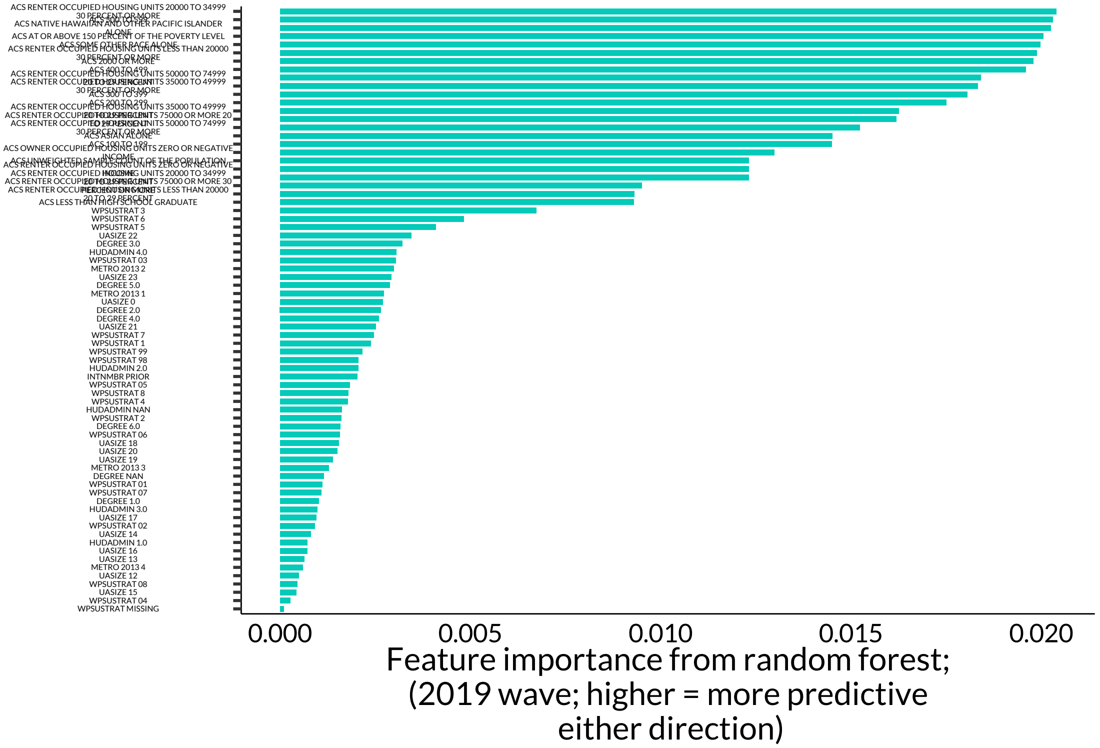


Table 10: Items in order effects analysis based on trace file The items are ordered by their average duration across respondents.

Variable	Average duration
MHMOVE	10.17
ENTRYSYS	11.83
HHSEX	12.81
GUTREHB	15.07
HOA	15.29
MHANCHOR	15.40
STORIES	15.89
STORIES_IUF	15.89
UNITFLOORS	15.89
TPARK	15.90
NOSTEP	16.21
HHMOVE	17.05
BEDROOMS	17.37
DINING	17.52
SOLAR	17.61
LIVING	17.70
KITCHENS	17.71
GARAGE	17.74
UNITSIZE	17.91
UNITSIZE_IUF	17.91
KITEXCLU	18.01
PORCH	18.03
WASHER	18.45
OTHFN	18.48
DENS	18.51
HHSPAN	18.57
FIREPLACE	18.75
LAUNDY	18.82
MONOXIDE	19.10
UFINROOMS	19.10
FRIDGE	19.15
COLD	19.20
KITCHSINK	19.28
SEWUSERS	19.34
HEATFUEL	19.69
FAMROOMS	19.74
NOWAT	19.74
WATSOURCE	19.81
COLDEQ	20.37
RECROOMS	20.42
VACMONTHS	20.51
NOWIRE	20.67
WALLCRACK	21.03
FLOORHOLE	21.15
TIMESHARE	21.15
FNDCRUMB	21.26
VACRESDAYS	21.26
ROOFSAG	21.27
WALLSIDE	21.31
ROOFSHIN	21.34
WALLSLOPE	21.37
COLDEQFREQ	21.41
ROOFHOLE	21.45
WINBOARD	21.49
MONLSTOCC	21.50
VACRNTDAYS	21.58
WINBROKE	21.60
WINBARS	21.63
NOWATFREQ	21.66
PLUGS	22.12
RENT	22.39
LOTVAL	22.49
YEARBUY	22.68
SUITYRRND	22.94
DWNPAYSRC	23.35
FIRSTHOME	23.96
FORSALE	24.25
LEADINSP	24.34
OILAMT	27.06
PROTAXAMT	27.73
TRASHAMT	28.28
WATERAMT	28.42
OTHERAMT	29.02
MOVWHY	34.15
RMJOB	34.32
RMOWNHH	34.48
RMCHANGE	34.61
RMCOMMUTE	34.63
RMFAMILY	34.65
RMHOME	34.82
RMCOSTS	34.98
RMHOOD	35.03
RMOTHER	35.08
SEARCHFAM	35.25
SEARCHNET	35.38
HHGRAD	35.42
NRATE	35.63
HHNATVTY	35.68
SEARCHPUB	35.68
HRATE	35.85
SEARCHOTH	35.86
SEARCHREA	35.88
SEARCHLIST	35.93
SEARCHSIGN	35.93
NEARWATER	36.53
HMRACCESS	37.47
NHQPCRIME	37.51
NHQSCHOOL	37.51
HMRENEFF	37.56
NHQSCRIME	37.69
HHINUSYR	37.70
HMRSALE	37.75
NHQPUBTRN	37.97
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 analyses

A.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.

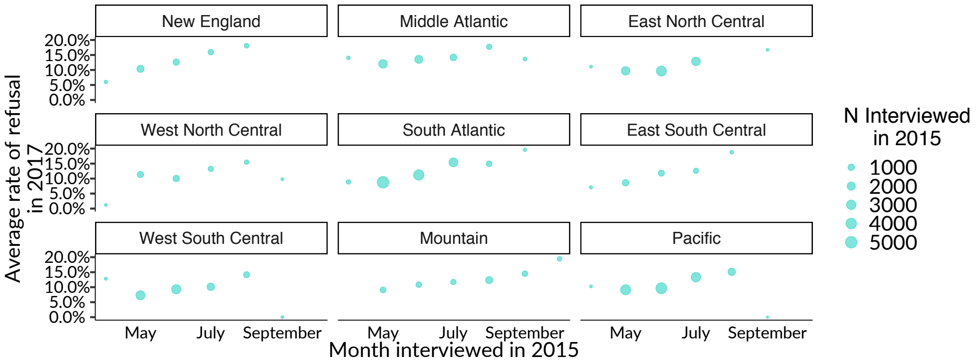


Table 12: Twenty-five next-strongest predictors of 2017 survey refusal among 2015 respondents.

term	estimate	std.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 non-refusers The proportions reweight using the composite weight.

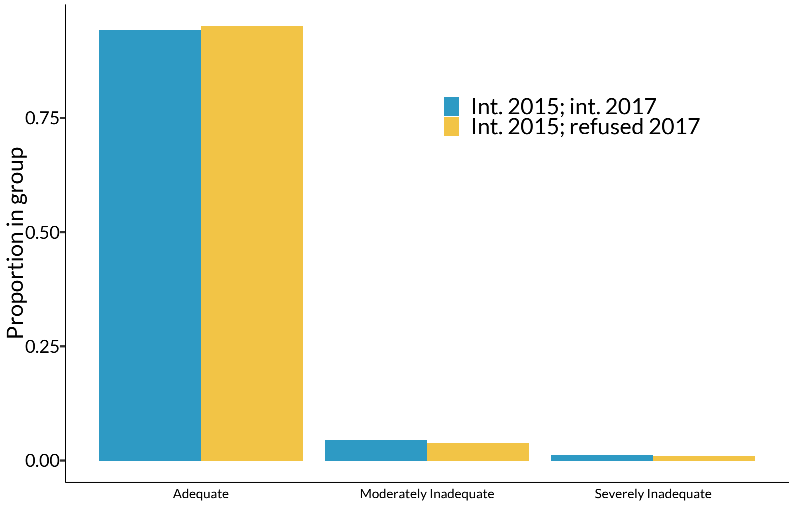


Table 13: Attritor heterogeneity in relationship between income and adequacy: refusers in 2017; regression. The table shows that in addition to main relationships where those with higher incomes are less likely to have inadequate housing, we see heterogeneity in this income-adequacy relationship between attritors and non-attritors.

	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*** (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