# Project Design

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

## 1 Project Objectives

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

As with many federal surveys, the AHS has experienced declining response rates, requiring increasing amounts of time and effort to reach the 80 percent response rate preferred by the Office of Management and Budget. In particular, response rates have declined from approximately 85 percent in the 2015 wave to 80.4 percent in the 2017 wave to 73.3 percent in the 2019 wave.[1]

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

In this project, we distinguish between nonresponse bias, on the one hand, and survey representativeness, on the other hand. Nonresponse bias is a divergence between a population quantity of key interest—such as the true proportion of U.S. adults living in severely inadequate housing—and its sample estimate, which arises due to systematic differences between those who do and do not respond to a survey.[2] In theory, it is possible to adjust survey estimates to account for differential nonresponse so that sample estimates converge to population quantities, and bias is removed. To account for potential nonresponse bias, the AHS calculates a nonresponse adjustment factor (NRAF) that reweights for nonresponse within cells defined by metropolitan area, type of housing unit, block group median income, and area-level rural/urban status. In principle, adjustments such as this, along with raking, should reduce or even remove the inferential threats posed by nonresponse bias. However, there is no guarantee that the model used for bias-adjusted estimates contains all the information it needs. Moreover, the weights used in such bias adjustment schemes typically increase variance in estimates: they essentially require units in grid cells with a lot of missingness to "represent" more unobserved units than those in grid cells with less missingness.

[1] 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.

[2] In other words, it is a correlation between the propensity to respond to the survey and a key outcome of interest.

Furthermore, our preliminary analyses leave open the possibility that the raking and nonresponse adjustment factors currently employed to reweight AHS estimates do not ensure convergence with population quantities. For example, a key outcome the AHS measures is housing inadequacy. Among units where an interview was successfully conducted during the 2015 wave of the AHS, some dropped out due to nonresponse in 2017. Reweighted estimates suggest 12 percent of those who stayed in the panel in 2015 and 2017 had problems with rodents. Looking at those housing units that appeared in 2015 only to drop out in 2017, however, only 9 percent had problems with rodents—in other words, a key measure of housing quality appears correlated with differential panel attrition. In a separate memo on nonresponse bias in prior rounds of the AHS (see attached), we found numerous systematic patterns in panel attrition whose statistical and substantive significance persists in spite of weighting meant to account for nonresponse bias. We found the AHS bias-adjusted estimate of the proportion of householders in the U.S. who own their home outright (without a mortgage or loan) in 2015 is seven percentage points lower than the corresponding proportion in the 2010 Decennial census count.[3] Attributing such divergence to nonresponse bias with complete certainty is a challenging task since, by definition, we cannot measure the outcomes of those who do not respond. However, the many pieces of evidence presented in the nonresponse bias memo suggest that, in addition to adjusting sample estimates on the backend, improving sample composition on the frontend would increase their accuracy.

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

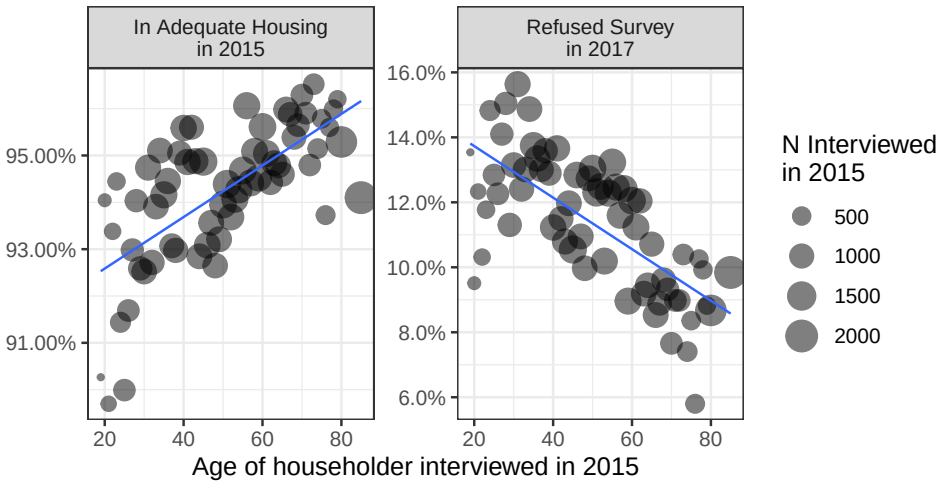


Figure 1: Units with young householders in 2015 were a) less likely to be adequate housing in 2015 and b) more likely to drop out of the panel due to refusal in 2017. Points represent reweighted estimates of proportions for different ages, size corresponds to number of respondents in 2015.

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

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

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

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

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

## 2 Intervention Design

Our intervention consists of sending cash to potential respondents sampled as part of the Integrated National Sample of the 2021 American Housing Survey. The cash is delivered inside an envelope containing a letter reminding the potential respondent about the survey. This letter is sent both to treatment and to control respondents, albeit with a slight wording change that mentions the incentive in the treatment letter and not in the control. The timeline is depicted on Figure 2.

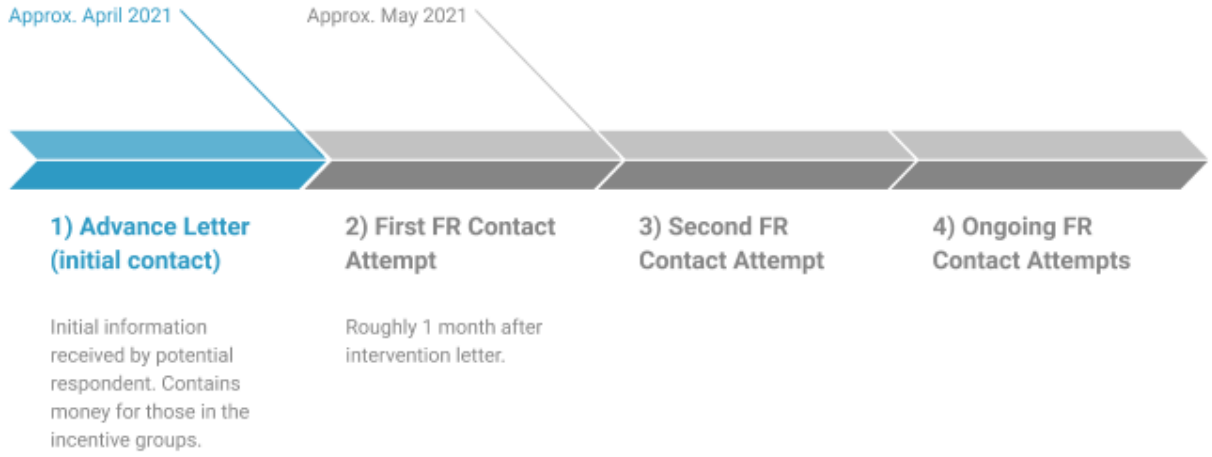


Figure 2: Intervention timeline.

Given the risks survey nonresponse raises—sample size reduction and possible bias—it is not surprising that a large literature has developed seeking to understand and reduce nonresponse. This project builds on a branch of this literature demonstrating the effectiveness of cash incentives at increasing response rates. We focus here on "noncontigent" and "nondiscretionary" cash incentives (Jackson, McPhee, and Lavrakas 2020). The cash incentives are noncontingent because they are provided to respondents in advance of the survey rather than only provided upon survey completion.[4] Second, the presence and magnitude of the incentive is nondiscretionary because it is determined centrally for all survey respondents, rather than at the discretion of individual field staff for particular respondents.

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

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

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

### 2.1 What contributes to survey nonresponse?

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

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

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

## 2.2 To whom should surveyors provide incentives?

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

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

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

While this procedure allows researchers to assess the impact of different incentive magnitudes, it ignores the fact that households differ in three ways. First is the household's likelihood of nonresponse. Second, among the pool of households with a low likelihood of response, is the extent to which that household's nonresponse contributes to bias. Third, among the pool of households with both a low likelihood of response and a high potential for that nonresponse to contribute to bias, is the extent to which that household is likely to be impacted by incentives. A growing set of literature seeks to: (1) identify these three groups, and (2) test approaches that target incentives on the basis of group membership.

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

Selecting respondents with the highest importance scores, the researchers randomly allocated the magnitude of incentive promised to survey respondents if they completed the survey (contingent incentive).[5] The study introduces an important conceptual approach to targeting—first, that incentives can be targeted to a subset of respondents and second, that researchers should take into account both response propensities and contributions to bias when selecting that subset. However, by giving incentives to *all* high importance respondents, it does not causally test whether targeting represents an improvement over randomly allocated incentives—the use of targeting as such is not evaluated. Similarly, other studies investigate different ways of operationalizing whom to target with incentives—for instance, Link and Burks (2013) compare response propensities estimated using different types of variables available in address-based sampling; Coffey and Zotti (2015) combine response propensities with sampling weights to find "highly influential" cases—but do not experimentally compare the effectiveness of targeting to the effectiveness of randomly-provided incentives. Such a comparison is crucial, however, in evaluating the effectiveness of targeting.

The most similar approach to ours is Jackson, McPhee, and Lavrakas (2020), which estimates response propensities and uses these to target incentives to complete a screener for the National Household Education Survey (NHES).[6] As in our proposed design, Jackson, McPhee, and Lavrakas (2020) randomly divides potential respondents into a group that

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

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

receives incentives independent of their propensity or one in which propensities determine incentive receipt. Specifically, the conditions are:

1. For the group assigned to propensity-independent incentives, respondents randomly receive either a $2 noncontingent incentive or a $5 noncontingent incentive along with their screener;
2. For the group assigned to targeted incentives, low propensity cases received $10, medium propensity cases received $5, medium-high propensity cases received $2, and very high propensity cases received $0.

Jackson, McPhee, and Lavrakas (2020) represents an important step forward for research on targeted incentives. However, its design has a fundamental drawback: the only group in which respondents receive no incentives is the targeted group. Thus, the effect of targeting is confounded with the effect of receiving no incentives. Unsurprisingly, giving high-propensity respondents $0 (in group 2) versus $2 or $5 (in group 1) decreases the response rate substantially. Thus, the study does not provide a good test of the targeting mechanism per se because it confounds targeting with the lack of incentives. Furthermore, predicting response based on demographic variables alone is notoriously difficult. Because the AHS is a panel survey of housing units, we are able to take advantage of a richer set of longitudinal data to improve the quality of the predictive models. In particular, in addition to the sampling frame data, we are able to include prior wave response outcomes (i.e., whether or not the unit responded) and prior wave paradata, which include the number, type, and timing of contact attempts and reasons for refusing the survey. We additionally leverage time-varying neighborhood characteristics from respective American Community Survey (ACS) 5-year estimates (2014; 2016; 2018) that capture aggregate demographic characteristics (age; employment status) potentially related to nonresponse. These data sources lead to a high-dimensional dataset with 400+ predictors; we use machine learning classifiers to retain this high-dimensional predictor set in the predictions of nonresponse.[7]

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

### 2.3 What is the right incentive amount?

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

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

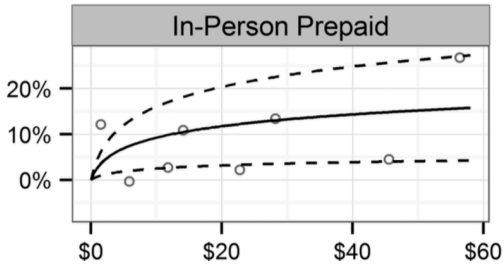


Figure 3: Dose-response effect of incentives on response rate in in-person surveys using noncontingent incentives, reproduced from the Mercer et al. (2015) meta-analysis.

Our study plans to randomize respondents to one of four amounts: $0, $2, $5, and $10. The $5 dollar amount is chosen as it corresponds to amounts in similar surveys such as the NHES. Figure 4 demonstrates examples of the response curves we might find. The dotted curves illustrate unobservable dose-response curves, while the solid curves and points show estimable quantities that the design can elicit.

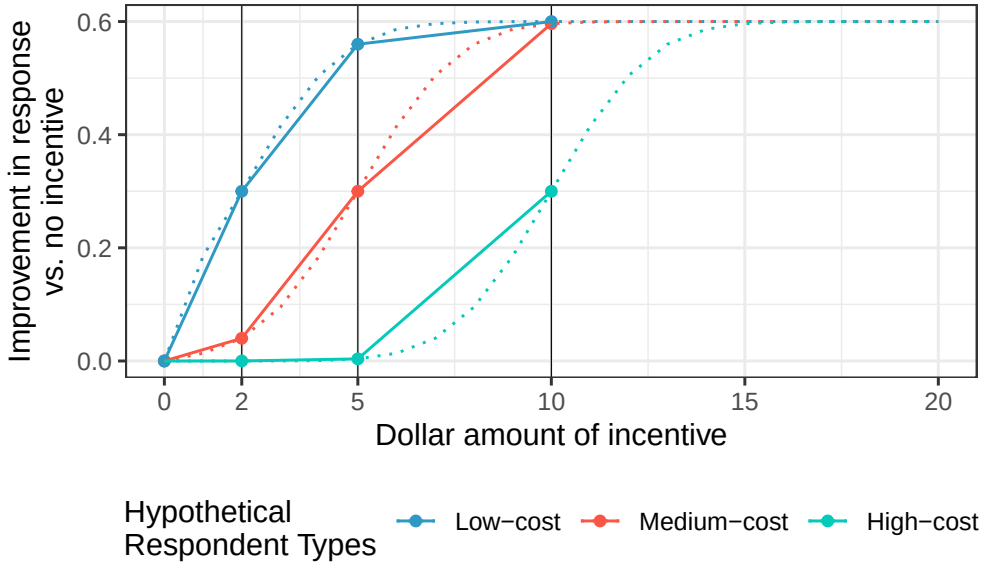


Figure 4: Possible dose-response curves and estimable linear relationships in the proposed study.

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

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

This raises the prospect that, either through habit-formation or recall, large incentive amounts may durably increase response rates beyond the one wave in which they are conducted or lead to an expectation of similar incentives in future waves. This is a possibility largely unexplored in the literature.

### 2.4 Remaining gaps in the literature

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

First, despite recognition that increasing the response rate overall does not necessarily reduce nonresponse bias (Groves 2006), studies largely continue to focus on response rates as the outcome to improve rather than measures of bias. In a meta-analysis published seven years after the points made by Groves (2006), Singer and Ye (2013) note an ongoing lack of research into the ability of incentives to address nonresponse bias. Since then, both Crissey, Christopher, and Socha (2015) and Jackson, McPhee, and Lavrakas (2020) target measures of bias as outcomes in addition to response rates, benchmarking characteristics of respondents to known quantities, but their studies have yielded no discernible improvements in bias from targeting.[8] Our study continues their work in examining reductions in bias, rather than improvements in response rates, as the primary outcome.

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

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

"An important outstanding question is whether it is possible to classify cases based not only on their base response propensity but also on the increase in response propensity that would be attributable to (for example) a higher incentive. If cases are heterogeneous in their sensitivity to an intervention, and if this sensitivity can be predicted from auxiliary data available prior to collection, then it may be efficient to target the intervention based on predicted sensitivity" (407).

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

### 2.5 Intervention Design

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

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

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

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

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

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

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

- some potential respondents will respond and others will not, so that the expected response rate is not 100% ($E[\bar{Y}] \neq 1$), which increases uncertainty by increasing the variance of the sample mean estimate ($E[\hat{\bar{X}}^2] - E[\hat{\bar{X}}]^2$),
- respondents will have different covariate profiles than non-respondents, with nonresponse bias defined as $\bar{X} - E[\hat{\bar{X}}]$. In general, we expect covariates to differ between people who respond and those who do not (for example, responders may be older, on average, than non-responders).

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

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

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

response rate can be reliably calculated, one might set $m = (1 - E[\bar{Y}])n$, so that the proportion of the sample that receives incentives, $m/n$, is equal to the proportion expected to not respond.

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

It also raises the question, addressed only imperfectly in the literature described above, of how that subset should be chosen. If it were possible to glean information on propensities to respond, would allocating incentives to those with the lowest $\eta_i$ increase response?[9] In addition, if the targeting is to those with the lowest propensity to respond, is there a subset of these low-propensity individuals who are most likely to introduce bias if not incentivized—that is, individuals that have attributes of interest that differ from those with high response propensities? How large would the gains from such an approach be?

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

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

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

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

## 3 Evaluation Design

We are interested in understanding how propensity-determined allocation of incentives affects nonresponse bias and how the size of incentives delivered to a potential respondent affects the rate of response among different subgroups

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

in the sample. The randomization is designed to generate the counterfactuals necessary to make these quantities estimable. We define these counterfactuals below.

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

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

Second, we denote by $Z^{T=1}$ the allocation of incentives that would have obtained had **Propensity-Determined Allocation** been used for the entire sample. The potential respondents are sorted by their $\hat{\eta_i}$, from lowest to highest, and an $n$-length vector of $m$ 1s and $n - m$ 0s is generated (with the 1s at the top and the 0s at the bottom.) Thus, those $m$ respondents with the lowest 30 percent of estimated propensities are guaranteed to receive an incentive, and those $n - m$ respondents with the highest 70 percent of estimated propensities are guaranteed not to receive an incentive. Note that this represents a considerable advantage: with an expected response rate of 74 percent, if our model does well at predicting nonresponse, we would be targeting all predicted nonrespondents—including both those on the margin and those who are perhaps less likely to be converted to responses—but a minimum of those already likely to respond.

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

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

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

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

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

As it turns out, however, this is a simple case of heterogeneous assignment probabilities. And, as we show below, it is easily dealt with using an inverse propensity weighted estimator. Specifically, since $T$ is independent, for any given individual the probability of assignment is given by Pr$(Z_i = 1) =$ Pr$(T_i = 1) \times Pr(Z_i = 1 \mid T_i = 1) +$ Pr$(T_i =$

$0) \times Pr(Z_i = 1 \mid T_i = 0)$. For the 30% ($m/n$) of units with the lowest propensity to respond (who will be allocated an incentive under targeting), this evaluates to $.5 \times 1 + .5 \times .3 = .65$. For the 70% of units with the highest propensity to respond (who will not be allocated an incentive under targeting), this evaluates to $.5 \times 0 + .5 \times .3 = .15$. Thus, there are four possible values of a treatment assignment probability $\pi^Z_{i,z}$ (where $z$ indicates an *observed* treatment status): for $j$ low propensity individuals, $\pi^Z_{j,1} = .65$ and $\pi^Z_{j,0} = 1 - .65 = .35$; for $k$ high propensity individuals, $\pi^Z_{k,1} = .15$ and $\pi^Z_{k,0} = 1 - .15 = .85$. Thus, it is possible to observe every unit in every treatment condition, albeit with differing probabilities. To obtain unbiased estimates of the average treatment effect, we simply downweight those who are overrepresented in treatment or control, and upweight those who are underrepresented, using $1/\pi^Z_{i,z}$, the inverse treatment propensity.

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

Finally, while the variation in $T$ that generates variation in $Z$ is the main variation we are interested in, we are also interested in the elasticity of incentives to response: $\partial \eta_i / \partial b_i$ and $\partial^2 \eta_i / \partial^2 b_i$. Thus, among those $m$ assigned to incentives, we plan to vary the amount of the incentive between 2, 5, or 10 dollars. This enables us to study the change in $\eta_i$ induced by a one-unit change in $b_i$. As we describe in greater detail below, this dose-response function could be highly non-linear. However, we are able to recover an estimand that is defined as a linear transformation of the potential outcomes using a linear estimator, even though the potential outcomes are generated through a non-linear process. See Figure 4 above for a graphical illustration of this point.

### 3.1 Total Number of Observations

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

### 3.2 Randomization / Assignment

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

2. **Create** $Z_i^{T=1}$. Order each potential respondent from highest to lowest $\hat{\eta}_i$. Calculate $m \approx .3 \times n$, and assign the first $m - n$ individuals to $Z_i^{T=1} = 0$ and the last $m$ to $Z_i^{T=1} = 1$. This provides the vector $Z^{T=1}$: the assignment that would have obtained, had each unit been assigned using Propensity-Determined Allocation.
3. **Create** $Z_i^{T=0}$. Define $f()$ as a function that randomly sorts a vector, and set $Z_i^{T=0} = f(Z_i^{T=1})$. This provides the vector $Z^{T=0}$: it is the assignment that would have obtained, had each unit been assigned to incentives using Propensity-Independent Allocation.
4. **Create** $T_i$. Sort individuals in order of their estimated propensity (randomly resorting within equal propensities) and form them into consecutive pairs. Within each pair, assign one individual to $T_i = 1$ and one to $T_i = 0$ with .5 probability. If there is an odd number of individuals, randomize the last unit using a coin flip.
5. **Create** $Z_i$. For all units for whom $T_i = 1$, set $Z_i = Z_i^{T=1}$, and for those for whom $T_i = 0$, set $Z_i = Z_i^{T=0}$.
6. **Create** $A_i$. Among units where $Z_i = 1$, randomly assign 50% to $A_i = 10$, 25% to $A_i = 5$, and 25% to $A_i = 2$. Assign the remaining sample for whom $Z_i = 0$ to $A_i = 0$.

### 3.3 Treatment Conditions

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

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

### 3.4 Outcomes

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

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

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

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

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

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

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

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

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

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

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

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

- Interpretation: This is the average effect of propensity-determined allocation on the mean of some covariate. This can be thought of as a more direct estimate of bias in the sense that we are able to directly observe estimates of key outcomes of interest for both treatment contitions. If propensity-determined allocation changes the proportion of groups likely to introduce bias above a propensity-independent allocation, we should be able to estimate this increase.
- Definition of estimand: $E[\hat{\hat{X}} \mid T_i = 1] - E[\hat{\hat{X}} \mid T_i = 0]$.
- How we estimate it: We regress $X_i$ on $T_i$. In the design diagnosis below, we refer to this estimand as the "Effect of T on sample mean(X)".

**Secondary Outcome: Effect of incentives on number of contact attempts**

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

### 3.5 Meaningful Effect Size

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

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

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

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

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

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

### 3.6 Likely Effect Size

Singer et al. (1999) compared the results of 39 experiments on financial incentives in face-to-face and telephone surveys. The effect sizes were smaller (though not statistically significantly so) for face-to-face surveys, translating to a one-third percentage point increase in response rates for each dollar spent. The treatment group will receive $2 \times .25 + 5 \times .25 + 10 \times .5 = 6.75$ USD on average, implying a likely average effect of $.003 \times 6.75 = .02 = \overline{\tau}$. In terms of average effects on if-untreated non-responders, this implies a 10 percentage point increase ($\tau = .10$). These parameters are assumed in the power calculations below.

### 3.7 Power

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

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

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

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

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

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

### 3.8 Data

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

### 3.9 Anticipated Limitations

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

1. **Our propensity model may not be good**. The design assumes that we are able to estimate $\eta_i$ in a reasonably informative way. If we don't have good propensity estimates, then any allocation of incentives on the basis of such estimates will be weaker. However, we are in a very favorable context in this study: we have panel data that has two years' worth of information about how respondents behaved in the past, as well as tract-level demographic information from the American Community Survey.

- *How we address*: In section 3 of the summary memo, we show that we can predict both nonresponse and refusal with a very high degree of accuracy in the 2017 and 2019 AHS. The most important predictors are past behavior—e.g., if a unit was a refuser in 2017 they are significantly more likely to continue to refuse in 2019. However, area-level demographics were also important predictors. We will use these findings to improve our ability to estimate $\eta_i$ in the targeting experiment.

2. **Developing estimates of $X$ from AHS data will be complicated**. The AHS data is not a simple random sample – data needs to be re-weighted to account for the sampling procedure in order to generate estimates. And our data also needs to be weighted by the inverse of the assignment propensities. So there is some complication here that is something of a risk – we need to make sure we get the weights right in order to say something meaningful about representativeness.

- *How we address*: in the summary memo, we discuss two considerations when comparing the AHS to benchmark data (in that case, the Decennial census). First is making sure that we align the variable definitions in each of the samples, which includes ensuring that we compare households to other households and that we compare questions asked in similar ways. Second is reweighting to account for the complex survey design. In the memo, and in follow-up discussions on proper weighting methods, we believe we can generate estimates of $X$ to properly compare to a benchmark population, both at the national and the CBSA level.

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

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

4. **Spillovers due to stopping rule**. The Census Bureau typically stops data collection once the target of an 80% response rate has been met. This poses a spillover concern for us: if we increase the response rate in area 1, then we may also decrease it in area 2 by reducing the need to collect more data there in order to achieve an 80% response rate.

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

## References


Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 99 (1): 544–55. https://doi.org/10.1257/aer.99.1.544.

Armstrong, J. S. 1975. "Monetary Incentives in Mail Surveys." *Public Opinion Quarterly* 39 (1): 111–16.

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

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

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

Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70 (5): 646–75. https://doi.org/10.1093/poq/nfl033.

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

Hidi, Suzanne, and K. Ann Renninger. 2006. "The Four-Phase Model of Interest Development." *Educational Psychologist* 41 (2): 111–27. https://doi.org/10.1207/s15326985ep4102_4.

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

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

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

Mercer, Andrew, Andrew Caporaso, David Cantor, and Reanne Townsend. 2015. "How Much Gets You How Much? Monetary Incentives and Response Rates in Household Surveys." *Public Opinion Quarterly* 79 (1): 105–29. https://doi.org/10.1093/poq/nfu059.

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

Singer, Eleanor, and Cong Ye. 2013. "The Use and Effects of Incentives in Surveys." *The ANNALS of the American Academy of Political and Social Science* 645 (1): 112–41. https://doi.org/10.1177/0002716212458082.