Survey Research Methods (2022)
Vol. 16, No. 1, pp. 97-131
doi:10.18148/srm/2022.v16i1.7767
European Survey Research Association

An evaluation of the quality of interviewer and virtual observations and their value for potential nonresponse bias reduction.

Weijia Ren
Westat
Rockville, USA

Tom Krenzke
Westat
Rockville, USA

Brady T. West
Institute for Social Research
University of Michigan, USA

David Cantor
Westat
Rockville, USA

With the decline of survey response rates over the past decade, survey researchers need to gather useful auxiliary variables for all sampled units and reduce the potential for nonresponse bias through adaptive survey design or nonresponse weighting adjustments. One potential source of auxiliary information is interviewer observations of characteristics of sampled units. Compared with area-level characteristics, which researchers often have available, characteristics at the dwelling unit level may provide more information about survey variables of interest and result in weight adjustments that could potentially reduce bias further. These observations, however, may vary greatly among observers, and may lack the quality needed for survey data producers. To investigate the quality and usefulness of such observations, this study systematically assesses completeness, validity, observer variance/reliability, and predictive power for bias reduction in a national pilot study for both in-person interviewer observations and virtual observations. This paper sheds light on the dwelling unit characteristics that are harder to observe, differences among interviewer and virtual observations, the potential value added beyond area-level characteristics for adaptive survey design and nonresponse adjustments, and ways to improve the observations.

Keywords: interviewer observations; virtual observations; nonresponse bias; adaptive survey design; weighting adjustments

1 Introduction

One of the biggest issues facing survey researchers is declining response rates. Although low response rates do not necessarily indicate nonresponse bias (Groves & Peytcheva, 2008), unit nonresponse may introduce bias in survey estimates when the characteristics and perspectives of the nonrespondents systematically differ from those of the respondents. Brick and Tourangeau (2017) re-emphasize that achieving higher response rates can potentially help reduce the impact of nonresponse bias on estimates. However, with response rates continuing to decline, survey methodologists have placed much more focus on reducing potential nonresponse bias through adaptive survey design (ASD) and weighting adjustments. The benefits of nonresponse adjustments rely on covariates that are associated with response propensity in the survey and with the survey variables of interest (Kalton & Flores-Cervantes, 2003). The ASD process can help survey organizations identify likely shortfalls in the sample, address problems in achieving the desired response rate and reduce the potential for nonresponse bias in the measured portion of the sample. For example, Rosen et al. (2014) discussed targeting low propensity cases in an effort to reduce potential nonresponse bias. Chapman (2014) and Tourangeau, Brick, Lohr, and Li (2017) also discussed ways to target cases to reduce potential nonresponse bias. Beaumont, Bocci, and Haziza (2014) emphasized call prioritization to minimize the variance of a nonresponse-adjusted estimator.

The ASD process and the weighting adjustment process both rely heavily on auxiliary data sources for a given sample (Kreuter et al., 2010; Little, 1986; Smith, 2011). The covariates should be available for the entire sample (both the respondents and nonrespondents) so they usually come from sources outside the survey questionnaire, especially if the nonrespondents do not participate in any stage of the survey. Data sources for these covariates typically include area-level information from a population census or a large sample survey producing public-use datasets (e.g, the American Community Survey (ACS)), sampling frames (e.g., pop-

Contact information: Weijia Ren, Westat, 1600 Research Boulevard, Rockville, MD 20850 (E-Mail: weijiaren@westat.com)

ulation registries), paradata, administrative data (if record linkage is possible), or existing datasets from commercial sources containing neighborhood and housing unit characteristics (Schräpler, Schupp, & Wagner, 2010; West, Wagner, Hubbard, & Gu, 2015).

For the same characteristics, unit-level data typically have more predictive power than area-level data (Hidiroglou & You, 2016). For example, the employment status for a person is a stronger predictor of food expenditures than a county-level proportion of persons employed in the population. However, for an in-person household survey in the United States, auxiliary information at the dwelling unit (DU) level is likely sparse, unreliable, or nonexistent. Throughout this paper, the term "DU" refers to a house, an apartment, a mobile home, a group of rooms, or a single room that is occupied as separate living quarters, in which the occupants live and eat separately from any other persons in the building and have direct access from the outside of the building or through a common hall (U.S. Census Bureau, 2019). The term "household" refers to DUs that are occupied.

Interest has emerged regarding the value of interviewer observations that are collected about characteristics of neighborhoods, DU types, and circumstances of the sampled units (Kreuter, 2013; Plewis, Calderwood, & Mostafa, 2017). Such observations are recorded by interviewers during the data collection process and describe the characteristics of sampled units as well as neighborhood characteristics. To the extent that such interviewer observations (IOs) are correlated with the response propensity (i.e., probability to respond) and with survey variables of interest, they may be useful in reducing potential bias arising from nonresponse in a cross-sectional context (Kreuter et al., 2010). There has also been some recent interest in gathering similar data from virtual observations (VOs) from Google Street View (Vercruyssen & Loosveldt, 2017). The virtual observations can be collected without venturing into the field and can provide similar auxiliary information.

Through our research, we seek to investigate the following questions:

1. What interviewer and virtual observation items tend to be of the highest quality?

2. What interviewer and virtual observation items have predictive power for reducing potential bias due to nonresponse through adaptive survey design or weighting adjustments, after taking into account the auxiliary data that already exists?

We argue that observations on these types of characteristics could have value if the quality of the observations is reasonably high. In this study, we set out to assess the quality and explore the predictive power of IOs and VOs on characteristics of the sampled units, such as neighborhoods, DU types, and household characteristics. The aim is to provide valuable insights into the usefulness of IOs and VOs for ASD and weight adjustment strategies as well as areas for improving the quality of such observations. Data quality is measured multi-dimensionally in this study. To assess data quality, we look at four criteria, namely:

1. Completeness (less missing data)

2. Validity (data accurately captured)

3. Interviewer variance (a measure of quality reflecting variance among interviewers) and virtual observer reliability

4. Predictive power (associated with survey response indicators and survey variables)

To this end, we make use of the Food Acquisition and Purchasing Survey (FoodAPS) Pilot Study in which IOs and VOs were made on a variety of these characteristics.

Some background material presenting past evaluations of interviewer and virtual observations is given in Section 2. In Section 3, we describe the FoodAPS Pilot Study data, including our evaluation's objectives and methods. The results of the evaluations of the interviewer observation and virtual observation data are provided in Section 4. Section 4 also includes results of the evaluation of the value added by the observations and cost comparisons of the two methods. The evaluation adds a "data point" to the literature on the value added from observation data for ASD and weighting adjustments beyond the auxiliary variables already available. Lastly, we provide a concluding summary in Section 5.

2 Background: Past Evaluations of Interviewer and Virtual Observations

Much of this literature is in the context of observations made by interviewers with evaluations that touch upon the properties mentioned above (completeness, validity, interviewer variance/virtual observer reliability, and predictive power). A guideline of an evaluation of interviewer observations on such key properties is provided by Sinibaldi, Durrant, and Kreuter (2013). In Section 2.1, examples from the literature are presented to address concerns about the low quality and weak predictive ability of interviewer observations. With similar goals in mind, a brief introduction to virtual observations is given in Section 2.2.

2.1 Interviewer observations

Interviewer observations usually capture observable classification variables rather than key survey variables. The extent of their utility for reducing potential nonresponse bias will therefore depend on the associations between the interviewer observation variables and both the survey response indicators and key survey variables (Lynn, 2003). Although interviewer observations can relate to both response indicators and key survey variables, these observations are typically interviewer judgments and are potentially prone to measurement error (West, 2013a). West and Kreuter (2013) demonstrate significant interviewer variance (i.e., variance across

and within interviewers as well as between different characteristics that are being assessed) in the quality of interviewer observations of DU features. Although they identify predictors of observation quality, it is less clear why unexplained variance in quality remains across interviewers when adjusting for area- and interviewer-level characteristics. If interviewer observations are not of consistently high quality, they can be problematic, and analyses depending on these data can be misleading. West, Kreuter, and Trappmann (2014) discuss error rates and variance in observation quality among interviewers for the German Labor Market and Social Security (PASS) study that may limit the modest predictive power of the observations. These authors found limited ability of the observations to predict response propensity. In another example, Kreuter et al. (2010) studied three participating countries in the European Social Survey (ESS) with the most complete interviewer observations, and concluded that the observations did not demonstrate the predictive power needed for successful nonresponse adjustment. In this study, the outcomes were TV watching and two items about trust. This conclusion may therefore not apply to other surveys with different subject matter.

Despite the concerns noted above, there are some indications about the potential fitness-for-use of interviewer observations. In the study by Sinibaldi et al. (2013) using United Kingdom census data, the authors found a high level of validity in some observations, and the interviewer observations that were analyzed suffered from minimal measurement error, resembled true values, and were usable for further analysis. Sinibaldi, Trappmann, and Kreuter (2014), in their study of PASS data, discuss certain observations being a better choice than purchased commercial data for nonresponse adjustments, especially observation items designed for subpopulations. The authors were also concerned about the quality of the observations, and they note that if observations with unsatisfactory quality are used in nonresponse adjustments, this could inhibit the intended impact of the adjustments (West, 2013b). In general, the data quality issues and predictive power vary by observation items—some items are more complete than others, some have higher validity than others, some are more reliable than others, and some have stronger associations with response propensity and survey outcomes than others. Through our research, we seek to identify observation items that satisfy all these criteria.

2.2 Virtual observations

While moderately successful in a business survey (Giangrande, Brick, Morganstein, & Lewis, 2018), many issues with virtual observations can arise in the residential setting. Clarke, Ailshire, Melandez, Bader, and Morenoff (2010) report coverage issues with Google Street View, with higher coverage of streets in urban areas. The authors also mention the time lag between the current day and the date that the pictures were taken. More recently, Vercruyssen and Loosveldt (2017) evaluated Street View in conjunction with the ESS to attempt validations of interviewer observations and also to assess the strength of virtual observations in predicting response propensity. The authors used logistic regression models to predict nonresponse (three outcomes in separate models: contact, refusal, and response) with auxiliary data (litter, vandalism, condition of home, impediments to access). Street View and interviewer observations were similar and did better in predicting contact propensity versus predicting refusal or response propensity.

The assessment by Vercruyssen and Loosveldt (2017) was conducted in Belgium and encountered pitfalls similar to those mentioned in Clarke et al. (2010). Street View was not available in some areas. There were also difficulties in finding homes, and if found, difficulties in making detailed observations. The authors also cited the time lag issue and that the observations took longer than expected. They found that similar predictions of nonresponse were made between the Street View observations and the interviewer observations. Mooney et al. (2017) compared in-person and virtual observations for physical disorder measures in Detroit. In this work, trained observers recorded social or physical characteristics of street segments according to explicit rules. They found low item-level reliability between observation types. However, there was a similar spatial distribution of physical disorder in Detroit computed using the two techniques. Because the in-person observations were recorded specifically for neighborhood audits, the paper concluded that virtual observations required only three percent of the recording time that was required by the in-person observations. Thus virtual observations were found to be a viable and much less expensive alternative to in-person observation for assessing neighborhood conditions. This conclusion, however, would not apply for in-person surveys with a household screening questionnaire.

In general, the studies mentioned above point to quite a few concerns, challenges, and areas for improvements related to the quality of both interviewer and virtual observations, but sometimes the observation data is of sufficient quality to help with further analysis. The observations in some studies seem to relate specifically to the subject matter at hand. It is useful, therefore, to assess these observation methods using the four criteria mentioned in Section 1, especially in light of the growing need for auxiliary data to help reduce nonresponse bias.

3 Data and Methods

This section provides relevant background about the Food Acquisition and Purchasing Survey (FoodAPS) and its data (Section 3.1). In Section 3.2, details are provided about the observation data collected. The objective of our investigation of the quality and value added of the observations is ex-

plained in Section 3.3. The associated evaluation methods are described in Section 3.4.

3.1 FoodAPS Background

The first FoodAPS was conducted in 2012 as a nationally representative survey of U.S. households that collected unique and comprehensive data about household food purchases and acquisitions. A household screener was administered in person, followed by an in-person interview. Using a paper diary, household members provided detailed information for seven consecutive days about foods purchased or otherwise acquired for consumption at home and away from home, including foods acquired through food and nutrition assistance programs. The survey's key domains included Supplemental Nutrition Assistance Program (SNAP) households, low-income households not participating in SNAP, and higher-income households. The sample design was comprised of a three-stage probability sample. The sample design began with the selection of 50 primary sampling units (PSUs), where the PSUs were single counties or groups of counties. Eight (8) block groups (i.e., clusters of blocks within the same census tract that have the same first digits of their four-digit census block number) were then sampled per PSU, and a sample of DUs was selected within each sampled block group. The sample design included an oversample of SNAP households and non-SNAP low-income households, while higher-income households were selected at a lower rate. To oversample SNAP households, states were asked for address lists of households on SNAP, and SNAP households were selected from that list with a higher rate than for other domains. The screening questionnaire was administered to help subsample high-income households to meet sample size requirements by sampling domains. A 41.5 percent overall unit-weighted response rate resulted after applying the American Association for Public Opinion Research (AAPOR) Response Rate 3 computation, where about half of the overall unit nonresponse occurred at the screening interview. The adjustment for nonresponse bias in the 2012 FoodAPS required that auxiliary information be available for all sampled DUs.

The data for this research came from the 2016 FoodAPS Pilot Study sample. The Pilot Study was conducted in preparation for the second FoodAPS with the objective of comparing results using a web-based diary approach in 2016 to the results gathered from the paper diary in 2012. To help improve comparisons between 2012 and 2016, the 2016 household sample was selected from PSUs that were also included in 2012. That is, similar to the 2012 sample design but smaller in scale, the first stage of sampling for the pilot study included a subsample of 12 PSUs from the 50 PSUs selected for the 2012 FoodAPS. In the second stage, an average of 10 Secondary Sampling Units (SSUs) were selected per PSU. SSUs were block groups or combinations of adjacent block groups. As shown in Table 1, 2,552 DUs were selected. Among the 2,552 sampled households, 2,470 households were observed by an interviewer, while 82 households were not successfully observed. Among the observed households, there were 2,143 occupied dwelling units. A screener questionnaire was administered to classify households into the following key sampling domains:

Table 1
Sample sizes by data collection stage

Data Collection Stage	Sample Size
Released dwelling unit sample	2,552
Dwelling units with completed interviewer observations	2,470
Occupied dwelling units (households) with completed interviewer observations	2,143
Completed screener	827
Completed screener and selected for main survey	687
Completed initial interview	473
Completed final interview	430
Completed final interview with completed interviewer observations	421

1. SNAP households (of any income);
2. Non-SNAP Women, Infants and Children (WIC) households (of any income);
3. Non-SNAP and non-WIC households with income at or below 130 percent of the poverty guideline;
4. Non-SNAP and non-WIC households with income above 130 percent and at or below 185 percent of the poverty guideline; and
5. Non-SNAP and non-WIC households with income above 185 percent of the poverty guideline.

As in 2012, the sample design included an oversample of SNAP households, while higher-income households were selected at a lower rate compared to the population distribution. To oversample SNAP households, states were again asked for address lists of households on SNAP, and SNAP households were selected from that list with a higher rate than for other domains. A subsample was selected of high-income households due to cost constraints and to help result in the oversample of SNAP households. There were 827 households that completed the screener, and 687 households that completed the screener and were selected for the main survey. Among these households, 473 completed the initial interview and were requested to complete a food log diary for seven days. The pilot study resulted in 430 respondent households who completed a final interview, among them,

421 with completed interviewer observations. The oversample resulted in 34.9% of the sample being SNAP households, which is over three times the percentage in the population. The distributions of some interviewer observations, such as the DU type, neighborhood type, and number of children, were found to be significantly different between the SNAP households and non-SNAP households. However, we will not discuss these differences since these differences are as expected and not the focus of this study.

Survey base weights were created to account for differential selection probabilities in the pilot study. Final weights were created by calibrating the base weights to external population control totals from ACS (for subgroups defined by race/ethnicity, number of children, household size, whether someone 60 years old or older resides in the household, SNAP participation, and household income) to enhance the representativeness of the estimates. Replicate weights were created using the delete-one jackknife approach to account for complex sample design effects in variance estimation.

3.2 Observation data collection

In the pilot study, we collected interviewer observations and virtual observations because the auxiliary data were generally very limited at the DU level. The interviewer observations were collected prior to the screening interview. A total of 56 interviewers were involved in the interviewer observation process, and the observed number of DUs ranged from as low as 1 DU to 130 DUs per interviewer. The observations were unobtrusive and required no interaction with members of the sampled units (Olson, 2013). An instrument for interviewer observations was developed for the pilot study (as shown in Appendix B). The interviewer observation form listed a total of eight closed-ended questions, including the DU type, the neighborhood type, whether there is evidence of a child/children living in the DU, estimation of the number of people living in the DU, whether the DU is well-kept, whether the house exterior is not maintained well, whether the DU has an abandoned vehicle around, and whether there is long grass next to the DU, as well as one open-ended question for any other observations.

These questions were carefully reviewed and selected from previous interviewer observation instruments used for other studies based on their relevance to the measures of interest in the FoodAPS study. Interviewers were trained on how to report on the condition of the DU based solely on the physical DU and the neighborhood. The form could be completed on a phone or tablet to make it easier. Interviewers were asked to record information about the condition of the DU and the type of DU prior to approaching the sampled DU or making contact with anyone in the home in the initial visit. They were instructed to record their best guess based on as much information as they could observe and choose the "unknown" options for some items if they were not sure about the answers. If interviewers spoke to anyone before completing the interviewer observation form, they were instructed to not complete it for that residence. Also, no changes to entries were to be made after the initial observation. The system did not prevent an interviewer from completing the form after the screener. In practice, interviewers adhered to this rule. Only 23 of the 2,470 observations had time stamps that showed the form was completed after the screener was completed. Since there were legitimate reasons that this could have happened (e.g., recorded observations on paper before doing the screener) and because there were such a small number of occurrences, these observations were left in the dataset.

For the virtual observations, initially DUs were assigned to four virtual observers about a year and a half after the pilot study was conducted. The virtual observers received training on how to code cases with examples from Google Street View. They were also trained to use their best judgment for the observation and record any issues they noticed for that particular DU. About halfway through the assigned work, one of the observers was replaced due to work circumstances. We treated the two observers as one, making a total of four virtual observers. One-third of each of the four virtual observers' cases was randomly assigned to another observer. This overlap was used to calculate inter-coder reliability. Each observer was provided more than 800 DUs and asked to observe via Google Street View and record their observations on 16 items. A small number (6 items) of the 16 items are the same as some of the 8 items mentioned earlier in the interviewer observation instrument. Both instruments are provided in Appendix B.

3.3 Evaluation objective

The FoodAPS pilot study interviewer and virtual observations were evaluated with respect to the four criteria presented in Section 1. The purpose of the criteria is to determine if the interviewer or virtual observations are of high quality and useful for reducing potential nonresponse bias through adaptive survey design and weighting adjustments. A secondary objective was to analyze the costs of the interviewer and virtual observations.

We reiterate that we will use the following four criteria to evaluate quality:

- Completeness
- Validity
- Variability and/or reliability
- Predictive power

3.4 Evaluation methods

To evaluate the quality of the observations on these characteristics being recorded, the observation completeness

needs to be evaluated; that is, there should be a very low proportion of missing values for an observation to be of high quality. Secondly, the validity of the items needs to be investigated to ensure that the observations are capturing the intended characteristic. Lastly, to assess data quality, the variability and reliability of the observations need to be assessed. In terms of variability, we mean the susceptibility the observation has to the differences in recorded values across interviewers, given the subjectivity of the interviewers recording the observations. By reliability, we assess how consistent the virtual observations are among the virtual observers. While the observations may be of high quality in terms of these initial dimensions, they also need to be related to the survey outcomes and response indicators. If so, they will have predictive power for reduction of potential nonresponse bias through adaptive survey design and weighting adjustments. The same evaluation methods were used for both interviewer and virtual observations with one exception (criterion 3), where we assess the variability for interviewer observations and reliability for virtual observations.

Criterion 1: Completeness. For the first criterion (completeness), we investigated the missingness of each item. We treat observations with no missing values as having a perfect rate of completeness, observations with less than 4 percent of missingness as high completeness, observations with 4 to 10 percent of missingness as moderate completeness, and observations with more than 10 percent of missingness as low completeness. Observations with high or perfect completeness may indicate ease of observation and will provide enough data for further analysis. For some items, observers are allowed to choose "Don't Know" or "NA", which are not considered as missing for this criterion.

Criterion 2: Validity. For the second criterion (validity), previous studies have found that interviewer observations based on first impressions and intuitions can be prone to error (West, 2013a), and those errors will reduce the effectiveness of post-survey nonresponse adjustments that use interviewer observations as part of a covariate set that forms weighting classes (Lessler & Kalsbeek, 1992; West, 2013b). Therefore, it is important to check the validity of the interviewer and virtual observations before using them in weighting procedures. We evaluated validity by checking the associations between the interviewer and virtual observations with reported values of related items from the survey or neighborhood information from other publicly-available sources. Specifically, there were four variables from the screener survey and two variables from ACS data (see details in Table C2). The reported survey data and ACS data are treated as "true values" in this analysis, although they may also be prone to measurement error. For example, if the interviewer observation indicates a high income household, and the household later reports a high income, then the observation is consistent with the actual reporting; otherwise, the observation is considered to be inconsistent or inaccurate.

We fitted a series of survey-weighted logistic regression models to the data, using the variables from the screener survey and ACS as the outcomes and the interviewer/virtual observation items as the predictors, with jackknife replicate weights to account for the complex sample design when estimating standard errors. In these models, the interviewer observation items were dichotomized and virtual observation items were categorized into three levels (i.e., "no", "yes", "missing/NA/DK"). In addition to the regression odds ratios, significance level, specificity and sensitivity from the fitted models, we also report the Kappa coefficients as indications of validity. Cohen's Kappa coefficient (κ) is usually used to measure the level of agreement, and the magnitude of κ indicates the strength of the agreement ($>$ 0.61 high, 0.41-0.6 moderate, 0.21-0.4 fair, and $<$ 0.2 low) (Landis & Koch, 1977).

Criterion 3: Interviewer Variance/Reliability. The third criterion differs between the interviewer and virtual observations. Interviewer observations are susceptible to the interviewers' use of subjective judgement, and this subjectivity is a source of variance that can be detrimental to their use in reducing potential nonresponse bias. In the pilot study, because it was not financially feasible to evaluate the inter-interviewer reliability during data collection, we used an ad-hoc method to explore the variability among the interviewers' observations. For example, as illustrated in Figure 1, we noticed that the proportion of interviewer-observed households with children did not vary as much across interviewers as the proportion of interviewer-observed households in high income neighborhoods. For illustration purpose, we restricted to interviewers who observed 30 DUs or more in the figure. Given the variation among interviewers, we analyzed the effects of the interviewers on the observations, adjusting for household characteristics from the screener questionnaire and neighborhood characteristics from the ACS data. The household characteristics may help to explain some, but not all, of the variation among interviewers. A limitation of this methodology is the nonrandom assignment of the sampled DUs to the interviewers.

Hierarchical generalized linear modeling (specifically, multilevel logistic regression) was used to investigate the interviewer variance (Raudenbush & Bryk, 2002), with the interviewer as a level-2 factor and households at level 1. The models were fitted in SAS v. 9.4 using PROC GLIMMIX (SAS Institute Inc., 2013); see Appendix A for more details.

To examine the reliability of the virtual observations, we randomly assigned the same DUs to a pair of virtual observers. Each DU was observed by two virtual observers, and we then computed agreement rates for each observation item. Also of interest was to compare the virtual observations to the interviewer observations where appropriate. Before doing this, we first "finalized" the virtual observations. When

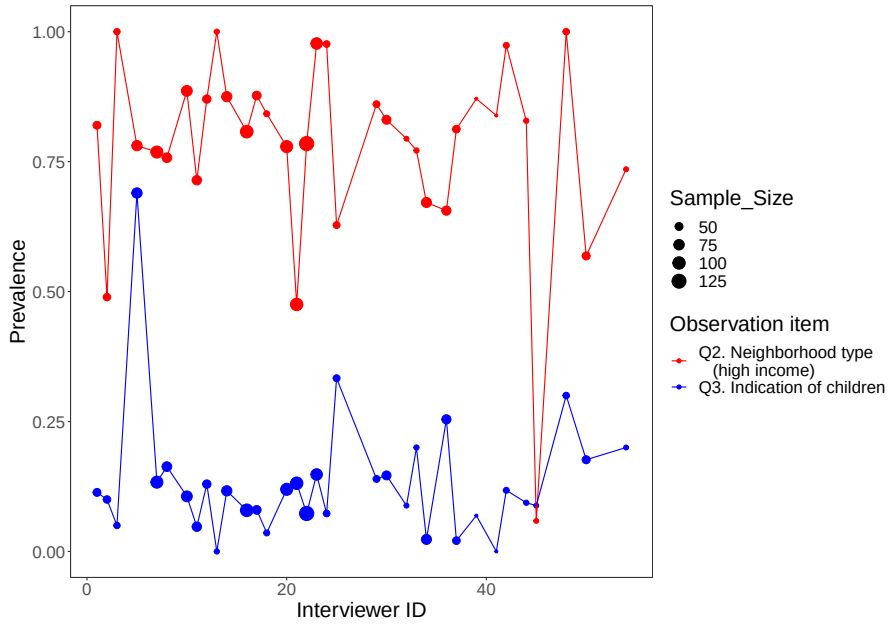

Figure 1. Proportions of households deemed to have children present or be in neighborhoods with high income, by interviewer (circles are proportionate to sample size)

the observations from the two virtual observers (VOers) were the same, we treated the matching observation as the "final" virtual observation. When the observations from the two VOers were different, we first ranked virtual observers by the agreement rate across all items, and then used the observations from the VOer with higher rank as the "final" virtual observation. Agreement rates were calculated as the weighted percentages of cases with agreement between the observers. We used the following rules of thumb to classify the agreement rates: above 99% is perfect, 95.1% to 99% is high, 90.1% to 95% is moderate, 75% to 90% is fair, and below 75% is low (Hartman, 1977).

Criterion 4: Predictive Power. We evaluated the fourth criterion (predictive power) based on the observations' associations with the response indicators and the key outcome variables. In order to use the observations to reduce potential nonresponse bias, the interviewer observations should be related to the response indicators and the key survey outcome variables (Little & Vartivarian, 2005). There is evidence that there exists a positive association between living in a DU in need of major repair and food insecurity (Kirkpatrick & Tarasuk, 2011). Specifically, in our study, the bivariate association between the observations and both the screener and final interview response indicators were analyzed, as well as with the five key survey outcomes: dichotomized food adequacy (0 = not enough food; 1 = enough food), number of food events/items acquired at home and away from home within a week (i.e., number of food-at-home (FAH) events, number of FAH items, number of food-away-from-home (FAFH) events, and number of FAFH items). Significant associations with comparatively higher effect size (i.e., odds ratio, or adjusted R-squared) may indicate favorable conditions for using the interviewer/virtual observations in ASD and weighting adjustments to reduce potential nonresponse bias. Linear and logistic regression models were fitted with the replicate weights. It should be noted that each regression model only contained one observation item as the predictor, given our objective to investigate the association of each observation item with the response indicators or key survey outcomes, without considering other observation items. We also noted that this analysis assumed all observation items were of good quality, so that they would provide maximum utility for nonresponse adjustment. Lack of quality may attenuate associations and limit the findings, as in West et al. (2014); in this study, however, we examine the predictive power of all items regardless of their quality, so that items with strong potential predictive power but low quality could be identified and possibly improved through training, thus being useful in future studies.

To further investigate the extent to which there is "value added" for either type of observation (interviewer vs. virtual), response propensity models were estimated by adjusting for auxiliary covariates[1] from the ACS or the screener questionnaire. The response propensity models were fitted in two phases. Phase 1 was conducted to reduce the set of auxiliary covariates. Four response propensity models were fitted: 1) a screener response propensity model used for ASD, 2) a

[1]For example, the proportion of the population with age 25 and up and less than high school education in the area.

screener response propensity model used for weighting, 3) a final interview response propensity model for ASD, and 4) a final interview response propensity model for weighting. In ASD models (model 1 and 3), information from the sample monitoring paradata (e.g., most recent interim response indicator before final status, number of screener contact attempts before final status) were also added to the models in addition to the area-level variables. For the weight adjustment models, the covariates included in the screener response propensity model (model 2) were area-level variables only. Both the area-level variables and screener variables (i.e., reported household size, children present, etc.) were considered for the final interview response propensity models (model 4), however, none of the screener variables were found to be significant, and thus only area-level variables were included in the Phase 2 models. Response propensity models for the screener and final interview were fitted under the context of improving ASD and the weight adjustment process.

Phase 2 evaluated whether any of the interviewer or virtual observation items improved the models created in Phase 1. Model selection was conducted via stepwise regression and by evaluating pseudo R-squared values. The following approaches were used to avoid potential multicollinearity among interview observation (IO) and virtual observation (VO) items when determining the final value-added IO and VO items:

1. Select IO items that add value to the model. Keep the selected IO items, and select VO items that add value to the model.

2. Select VO items that add value to the model. Keep the selected VO items, and select IO items that add value to the model.

3. Use the IO items and VO items selected in the first two steps to create the final model.

4 Results

We present the results of the interviewer observation and virtual observation evaluations in Section 4.1 and Section 4.2, respectively. Section 4.3 provides the results from fitting the value-added models, and the cost comparison results are explained in Section 4.4.

4.1 Interviewer observation results

This section systematically presents the evaluation results for each of the four key criteria, and concludes with a summary.

Criterion 1: Completeness. Among the 2,552 sampled DUs in the pilot study, interviewer observations were collected for 2,470 (96.8%) DUs. Interviewer observations were attempted on both vacant and occupied DUs (there were 2,143 occupied DUs with completed interviewer observations as shown in Table 1). Among these DUs, 421 households completed the final interview (another 9 households completed the final interview but did not have interviewer observations). All of the eight IO items had no item nonresponse. Among them, for two items, IO3 (indication of children) and IO4 (number of residents), interviewers had been encouraged to make their best guess, but it was also okay to indicate "Don't Know" as a valid response option. Interviewers selected the "Don't Know" option 14 percent of the time for IO3 and 20 percent of the time for IO4. Table C1 in Appendix C shows the frequency distributions of the eight IO items. Item IO1 (DU type) and IO2 (neighborhood type) are dichotomized for ease of analysis. Item IO5 (house well kept) is reverse-coded as "indication of house not well kept" so the estimation is in the same direction as the other items. Based on the low prevalence of unit and item nonresponse, we conclude that a high rate of completeness is achieved for the interviewer observations.

Criterion 2: Validity. The validity was tested using logistic regression models and the Kappa coefficients for all the IO items except IO5-IO8 due to lack of corresponding survey items. DUs with the "DK" option for IO3 (indication of children) or IO4 (number of residents) were not included in the validity analysis. As shown in Table C2, all the items are found to be significantly associated with the survey items or ACS variables. Among them, IO2 (neighborhood type) is found to have the strongest association, while IO4 (number of residents) has the weakest association, in terms of the magnitude of the odds ratios. The specificity and sensitivity also vary among the IO items. Some items (i.e., IO1 and IO2) have higher specificity while others (IO3, IO4) have higher sensitivity. It should be noted that the validity test using variables from the screener survey could only be conducted among responding households that responded to the related survey items, while the validity test using variables from ACS data could be conducted among all sampled households.

In addition to the regressions, the Kappa coefficients were also calculated. Among the four items, IO1 (DU type) and IO2 (neighborhood type) have higher Kappa coefficients (κ = 0.31 to 0.41), while IO3 (indication of children) and IO4 (number of residents) have lower Kappa coefficients (κ = 0.21 to 0.23). In general, the results of these analyses suggest that the interviewer observation validity is fair (0.21–0.4) according to the Kappa coefficients (Landis & Koch, 1977), where comparatively, items IO1 and IO2 have higher validity than items IO3 and IO4.

Criterion 3: Interviewer Variance. We fitted seven multilevel models, and the results are presented in Table C3 in Appendix C. We examined the estimated variance components and tested whether they were significantly greater than 0 according to a likelihood-ratio test based on a mix-

ture of chi-square distributions. Due to the small numbers of reported occurrences, models were not fitted for IO items related to abandoned vehicles or long grass. Significant variation across interviewers would be identified by having significant random variation for the intercept term (β_{0j}). Cases with the "Don't Know" option for IO3 and IO4 were excluded from the modeling.

The results show that there is evidence of significant interviewer variance for most of the characteristics examined. For example, for IO2 (neighborhood type), the variance component for the random intercept β_{0j} is 1.78, which is significantly greater than 0, indicating significant variation of IO2 across interviewers. Similarly, after adjusting for the available household and neighborhood characteristics, there exists evidence of unexplained interviewer variance for almost all the items except the more straightforward item IO1 (DU type). Given these findings, we conclude that more concrete guidance may be needed during training to stabilize the distributions of the observations across the interviewers.

Criterion 4: Predictive Power. The predictive power was tested by measuring the association of the IO items with the response indicators and key survey outcomes. Table C4 shows the results from the logistic regression models using each response indicator as the outcome and each interviewer observation item as the predictor. Ineligible households (i.e., vacant DUs, seasonal DUs, and DUs that are unable to locate and unable to enter) didn't have the response status and were excluded from these analyses. Item IO3 (indication of children) and IO6 (house exterior not maintained) are the only items related to response indicators for both the screener and final interviews. IO2 (neighborhood type) and IO5 (house not well kept) are each associated with the screener response indicator only. The remaining items are not associated with either response indicator. Items IO7 (abandoned vehicle) and IO8 (long grass) were observed for less than two percent of the sampled DUs, which is likely the reason why no significant associations are found.

The next consideration is the set of associations among the IO items and the key survey outcomes. Table C5 presents evidence of significant relationships between food adequacy and IO1 (DU type), IO2 (neighborhood type), and IO6 (house exterior is not maintained). It should be noted that only responding households who completed the final interview ($n = 421$) are included in these analyses. Table C6 presents the results in separate columns for the number of food events and items, separately for food-at-home (FAH) and food-away-from-home (FAFH) events. IO5 (house not well kept) is found to be a significant predictor for three out of the four measures. IO4 (number of residents) and IO8 (long grass) are significant predictors for two out of the four measures. IO2, IO3 (indications of children), IO6, and IO7 (abandoned vehicle) are significant predictors for only one of the four measures.

In summary, all interviewer observation items are found to be associated with at least one of the five outcome measures (food adequacy, FAH food events, FAH food items, FAFH food events, FAFH food items), with IO5 associated with the most measures (3 out of 5).

Interviewer Observation Quality Assessment: Summary. Table 2 summarizes the results based on the four criteria. All the IO items are reported to have perfect completeness, with a moderate rate of "Don't Know" responses for IO3 (14%) and IO4 (20%). Among the items that can be compared to survey items (IO1–IO4), the validity is fair, where IO1 and IO2 have higher validity than IO3 and IO4. There is evidence that most of the items exhibit significant interviewer variance after adjusting for household and neighborhood characteristics, with an exception for IO1. For IO7 and IO8, the models could not be estimated due to the small number of observed cases. Four items (IO2, IO3, IO5, IO6) are associated with at least one of the response indicators. All of the observation items (IO1–IO8), however, are associated with at least one of the five key survey measures. Item IO5 is associated with the most key survey measures (3 out of 5).

Across all the interviewer observation items, no one item satisfies all the evaluation criteria (completeness, validity, interviewer variance, and predictive power). For example, item IO1 is found to have comparatively higher quality based on completeness and interviewer variance, but lower in terms of validity and predictive power. On the other hand, item IO5 has comparatively higher predictive power but shows high variation among interviewers. Nevertheless, all items except IO7 and IO8 are possible candidates for further investigation toward use in ASD or weighting adjustment, or for use in a nonresponse bias analysis.

4.2 Virtual observation results

We investigated the FoodAPS pilot study virtual observations with respect to the four criteria in the same manner, and the results are discussed below.

Criterion 1: Completeness. Virtual observations were completed for 2,469 (96.7%) of the 2,552 sampled DUs. Table C7 gives a description of the virtual observation items with frequencies. There are a total of 16 virtual observation items (VO0–VO15). If the observers did not make a choice, it was treated as missing. Among the 16 items, most have missing rates of less than 4 percent except for VO2 (DU visibility, 6%), VO3 (DU type, 4%), and VO5 (neighborhood type, 7%). Observers were allowed to choose "Don't Know" (DK) if they were not sure about the answer for some items. The items a with high proportion of DKs are VO1 (DU number, 51%), VO6 (number of residents, 27%), VO7 (DU not well kept, 32%), VO10 (litter/vandalism, 29%), and VO12 (indication of children, 37%). The results are similar to, but not as high as the interviewer observation completeness, and suggest high completeness in general, but items such as the

Table 2
Summary of quality assessments for the interviewer observations

Interviewer Observation Item	Completeness[a]	Validation	Interviewer variance	Response indicator (associated / total possible)	Key survey measures (associated/ total possible)
IO1. Dwelling unit type	Perfect	Fair	Not significant	0/2	1/5
IO2. Neighborhood type	Perfect	Fair	Significant	1/2	2/5
IO3. Indication of children	Perfect (14% DK)	Fair	Significant	2/2	1/5
IO4. Estimated number of residents	Perfect (20% DK)	Fair	Significant	0/2	2/5
IO5. Indication of house not well kept	Perfect	N/A	Significant	1/2	3/5
IO6. Indication of house exterior not maintained	Perfect	N/A	Significant	2/2	2/5
IO7. Indication of abandoned vehicle	Perfect	N/A	N/A	0/2	1/5
IO8. Indication of long grass	Perfect	N/A	N/A	0/2	2/5

[a] IO entrees of "Don't Know" are treated as non-missing.

number of residents and indication of children are hard to observe. The items DU-not-well-kept and litter/vandalism have high DK rates in the virtual observations but not in the interviewer observations, which indicates possible advantages for in-person observation compared with virtual observations on such items. It also should be noted that in VO1, about half of the DU address numbers could not be visually confirmed due to the angle and clarity of the pictures, but sometimes the observers could infer the numbers from the nearby households.

Criterion 2: Validity. As shown in Table C8, the validity tests were evaluated via logistic regression and Kappa coefficient estimates. For the purpose of analysis, missing and Don't Know (DK) responses were combined in the "missing" category. Similar to the interviewer observation analysis, the validity test could only be conducted on four VO items (VO3, VO5, VO6, and VO12). Most of the test results are significant (in terms of the coefficient estimates) except for the association between VO12 (indication of children) and households having children under age 5, which is not significant. This association also has the lowest Kappa coefficient ($\kappa < 0.01$). The test between VO12 and households having children under age 18, on the other hand, shows a significant positive association with a low Kappa coefficient ($\kappa = 0.14$), which is higher than the association with children under age 5. The inconsistent significance results indicate that the associations between VO12 and the two variables (i.e., children under age 5 and children under age 18) are not consistently strong. One possible explanation could be that the Street View pictures of the DUs were not taken recently, so the lag between the time when the pictures were taken and when the observations were conducted might be long enough to misclassify households according to having young children or not (but may not affect identifying children under age 18). An advantage of the interviewer observation, in this case, is that the observers can see more details or hear sounds that the pictures do not capture. The item VO3 (DU type) is found to be the most valid item among the four items, in terms of the large odds ratio and highest Kappa coefficient ($\kappa = 0.38$). However, in general, the validity of the four items is low to fair.

Criterion 3: Reliability. To evaluate the reliability of the VO items, agreement rates among virtual observers themselves and between virtual observers and interviewers observing the same DUs were explored. High agreement rates indicate high reliability, and the rating scale follows the rule introduced in section 3.4. Table C9 shows that among virtual observers, item VO0 (Street View availability, 87%) and VO3 (DU type, 86%) have the highest agreement rates, whereas VO14 (rooftop width, 49%), VO6 (number of residents, 54%), and VO11 (number of windows, 56%) have the lowest agreement rates, revealing low observer reliability for these items.

Table C10 shows that between virtual observers and interviewers, item VO3 (DU type) has the highest agreement (84%), while VO6 (number of residents) has the lowest agreement (39%), which again confirms that the DU type is a comparatively more stable item among different observers, while number of residents could be highly unreliable. How-

ever, in general, the reliability among virtual observers or between virtual observers and interviewers is low, which indicates that more thorough trainings may be necessary with details and practical examples.

Criterion 4: Predictive Power. The virtual observation items were also evaluated for their associations with the response indicators and key survey outcome variables. Similar to the interviewer observations, significant associations would suggest possible benefits to use the VO items in ASD and weighting adjustments to reduce potential nonresponse bias.

Separate logistic regression models were fitted with the response indicators as the outcome (Yes vs. No) and each virtual observation item as the predictor. Each predictor had three categories (i.e., No, Yes, Missing/DK), and comparisons between the "Yes" and "No" categories, as well as the "Missing/DK" and "No" categories were tested. Table C11 presents the odds ratios and significance of the models. Most of the VO items were not associated with either response indicator. However, for VO5 (neighborhood type), DUs from middle/high income neighborhoods are significantly less likely to respond in the screener interview than those from low income neighborhoods (odds ratio = 0.68, $p = 0.0067$). It should be noted that even though there are no items significantly associated with the final response indicator, item VO8 (existence of sidewalk) has a marginally significant association (odds ratio = 1.28, $p = 0.0668$) where DUs with a sidewalk are more likely to respond in the final interview than DUs with no sidewalk.

The associations between the virtual observation items and the key survey outcomes (food adequacy, FAH/FAFH items/events) were also tested. The results in Table C12 reveal that only VO2 (DU visibility), VO6 (number of residents), and VO12 (indication of children) are found to be significantly associated with food adequacy. Specifically, DUs that are visible are significantly more likely to report that they have enough food to eat (odds ratio = 3.2, $p = 0.0464$). DUs for which the observers do not know the number of residents or not sure whether there are children present are significantly less likely to report that they have enough food to eat than DUs with one to two residents (odds ratio = 0.25, $p = 0.0101$) or DUs with no indication of children (odds ratio = 0.22, $p = 0.0030$).

For the number of FAH and FAFH food events and items, Table C13 provides the regression coefficient estimates and adjusted R-squared values from the models. Across all the models, the adjusted R-squared values are low (less than 0.04), indicating that the virtual observation items could only explain a small portion of the variation in the outcomes. However, some items are found to be significantly associated with some of the outcomes. Item VO10 (litter/vandalism) is significantly associated with all four outcomes, and evidence of litter/vandalism is associated with fewer reported food events and items. Item VO4 (gated community), VO13 (nearby non-residential buildings), and VO14 (rooftop width) are found to be associated with three out of the four outcomes, where DUs in a gated community, near a non-residential building, or with a shorter rooftop width tend to report fewer food events and items. Item VO5 (neighborhood type), VO6 (number of residents) and VO11 (number of windows in the front face of DU) are associated with two of the four outcomes, where DUs in low-income neighborhoods, with fewer people living in the DU, or with fewer windows tend to report fewer food events and items.

Virtual Observation Quality Assessment: Summary. The results from analyses of the virtual observation items with respect to the four major criteria are summarized in Table 3. The completeness is generally high across all the items except VO2, VO3 and VO5, but some items are found to have higher "Don't Know" rates than others, including VO1, VO6, VO7, VO10, and VO12. The findings are consistent with the interviewer observation results where the number of residents (VO6) and evidence of children (VO12) are hard to observe. In addition, the DU number (VO1), whether the DU is well-kept (VO7), and evidence of litter/vandalism (VO10) are also hard to observe from the Street View pictures due to the angle, scope and quality of the pictures.

The validation assessment can only be performed for item VO3, VO5, VO6, and VO12 due to the lack of corresponding survey items or ACS variables for the other VO items. In general, the virtual observation items have low-to-fair validity, among which VO3 and VO5 have higher validity than VO6 and VO12.

The observer reliability is low to fair across all items. Among the virtual observers, item V00 (Street View availability) and VO3 (DU type) are found to have the highest agreement, while VO6 (number of residents), VO11 (number of windows) and VO14 (rooftop width) are found to have the lowest agreement. Between virtual observers and interviewers, VO3 (DU type) and VO5 (neighborhood type) are found to have the highest agreement, while VO6 (number of residents) has the lowest agreement, which agrees with the interviewer observation results, and again indicates the difficulty of observing this characteristic.

In terms of associations with response indicators, VO5 (neighborhood type) is significantly associated with the screener response indicator, which aligns with the interviewer observation findings. VO8 (existence of sidewalk) is marginally associated with the final interview response indicator. For the key survey outcomes, VO10 (litter/vandalism) is associated with the most key survey outcomes (4 out of 5). VO4 (gated community), VO6 (number of resident), VO13 (nearby non-residential building), and VO14 (rooftop width) are found to be related to more than half of the key survey outcomes (3 out of 5).

Across all the virtual observations, not one single item

satisfies all the evaluation criteria. That being said, VO3 (DU type), VO4 (gated community), VO5 (neighborhood type), VO6 (number of residents), VO8 (existence of sidewalk), VO10 (litter/vandalism), VO12 (indication of children), VO13 (nearby non-residential building), and VO14 (rooftop width) are good candidates for further investigation based on high reliability and some evidence of association with the key survey outcomes.

4.3 Evaluation of value added

Through the two-phase variable selection process described in Section 3.4, different sets of items were retained in the four models (screener response ASD, screener response weighting, final interview response ASD, final interview response weighting). Table 4 summarizes the variables included in each of the four models.

The screener response propensity model for ASD is dominated by the interim interview response code (e.g., initial refusal). However, as shown in Table 5, two interviewer observation items [IO2 (neighborhood type) and IO3 (indication of children)] and one virtual observation item [VO9 (availability of driveway)] provide additional information for the screener ASD model. The pseudo R-squared value increases slightly from 0.348 to 0.351 with the inclusion of the IO and VO items. For the screener response propensity weighting model, two interviewer observation items [IO3 (indication of children) and IO4 (number of residents)] provide additional information for the screener weighting model; however, none of the VO items are significant. The pseudo R-squared value increases from 0.013 to 0.023 when including the two IO items.

For the final interview response propensity models that would be used for ASD and weighting adjustment, one virtual observation item [VO8 (existence of sidewalk)] is found to add additional value to the model, increasing the pseudo R-squared value slightly from 0.289 to 0.299 for the ASD model, and from 0.053 to 0.066 for the weighting model. None of the interviewer observation items are found to add value to the final interview response propensity models.

4.4 Evaluation of cost

The costs of the two observation approaches also play an important role in the evaluation of whether to use these observations in practice. As shown in Table 6, among the four virtual observers, the number of hours completing the task range from 35 to 64 hours. The average recording time per DU is 3.7 minutes. For the 2,552 sampled DUs, this task would take 157 hours. The staff who performed this task were entry-level statisticians. Suppose their pay rate is $50 per hour; it would cost $7,850 for the 15 virtual observations items on the 2,552 DUs. For interviewer observations, the time spent on the observation is considered as part of the interviewing process. If we assume that 2,552 observations were made, and it required 2 minutes per DU (based on internal estimates), it would take 85 hours total for the interviewer observations.

5 Conclusion

With decreasing response rates in surveys and increasing potential for nonresponse bias, it is important to find alternative data sources that could possibly reduce nonresponse bias. Interviewer observation data might be a useful source, and therefore we sought answers to the following questions:

1. What observation items result in high quality data?
2. What observation items result in predictive power for potentially reducing bias due to nonresponse through adaptive survey design or weighting adjustments, after taking into account the auxiliary data that already exist?

For all sampled DUs in the FoodAPS pilot study, data from eight interviewer observation items and 16 virtual observation items were collected. A systematic evaluation of both interviewer and virtual observation data was conducted according to four criteria, where the first three relate to data quality: 1) completeness, 2) validity, 3) interviewer variance/reliability, and the last one is related to the ability to predict response propensity and survey outcomes: 4) predictive power. Tables 7 and 8 provide a high-level summary of the evaluation results for each observation item.

5.1 Quality of observations

In terms of quality, the evaluation shows high rates of completeness among all the interviewer and virtual observation items, where interviewer observations have a higher measure of completeness than the virtual observations. For the tests conducted for validity, the interviewer and virtual observation items show indications of low-to-fair validity among the items that could be checked, where the DU type and the neighborhood type have relatively higher validity. Interviewer observations also show higher validity than the virtual observations in general. In terms of the interviewer variance, evidence exists of interviewer variance in the recorded observations, with the exception of DU type. For virtual observations, we found a wide range of low-to-fair estimated reliability (Street View availability and DU type are the most reliable).

In general, our conclusion is that the validity and reliability of the interviewer and virtual observations should be improved for most items before including them in ASD or weighting adjustments. This would require better training of the interviewers and observers on this process. In the case of the interviewer observations, neighborhood type could be considered as good quality except for interviewer variance; and emphasis on training would be helpful in limiting the subjectivity of interviewers' observations. The pilot study training was limited to emphasizing that observations should be made before they start the interview or interact

Table 3
Summary of quality assessment of virtual observations

Virtual Observation Item	Completeness	Validation	Reliability		Response indicator (associated/ total possible)	Key Survey outcome (associated/ total possible)
			Agreement among VOs	Agreement between VOs & IOs		
VO0. Was Street View available?	High	N/A	Fair	N/A	0/2	0/5
VO1. Can you visually confirm the dwelling unit number?	High (51% DK)	N/A	Low	N/A	0/2	0/5
VO2. Is the dwelling unit visible?	Moderate	N/A	Low	N/A	0/2	1/5
VO3. What type of dwelling unit?	Moderate	Fair	Fair	Fair	0/2	0/5
VO4. Is there a locked gate that impedes access to the dwelling unit?	High (15% DK)	N/A	Low	N/A	0/2	3/5
VO5. What type of neighborhood?	Moderate	Fair	Fair	Low	1/2	2/5
VO6. What is your best guess of the number of people living in the dwelling unit?	High (27% DK)	Low	Low	Low	0/2	3/5
VO7. Is there any indication that the dwelling unit is not-well-kept?	High (32% DK)	N/A	Low	Low[a]	0/2	0/5
VO8. Does a sidewalk exist in front of the dwelling unit?	High (8% DK)	N/A	Low	N/A	0/2	0/5
VO9. Is there a driveway or parking lots for the dwelling unit?	High (8% DK)	N/A	Low	N/A	0/2	0/5
VO10. Is there evidence of litter or vandalism?	High (29% DK)	N/A	Low	Low[b]	0/2	4/5
VO11. How many windows do you see in the front face of the dwelling unit?	High (20% DK)	N/A	Low	N/A	0/2	2/5
VO12. Is there any indication of a child or children living in the dwelling unit?	High (37% DK)	Low	Fair	Low	0/2	2/5
VO13. Full circle, do you see any non-residential buildings?	High (13% DK)	N/A	Low	N/A	0/2	3/5
VO14. How many car lengths is the rooftop width?	High (10% DK)	N/A	Low	N/A	0/2	3/5
VO15. Is the street condition good?	High (11% DK)	N/A	Low	N/A	0/2	0/5

[a] The agreement between VO7 and IO items are the average between VO7 & IO5, and VO7 & IO6. [b] The agreement between VO10 and IO items are the average between VO10 & IO7, and VO10 & IO8.

Table 4
Variables included in the screener and final interview response propensity models

	Screener response propensity model[a]		Final interview response propensity model[a]	
Covariantes	ASD	Weighting	ASD	Weighting
Sample monitoring paradata				
Most recent interim response indicator before final status (0=No, 1=Yes)	✓	-	✓	-
Number of screener contact attempts before final status	✓	-	✓	-
Area-level variables				
Proportion age 25+ with less than a high school education	✓	✓	✓	✓
Proportion non-Hispanic Black alone	✓	✓	✓	✓
Completed 2010 Census mail forms received from addresses in a mailback type of enumeration area (Mailout/Mailback and Update/Leave areas) out of all addresses from which a Census form was expected to be delivered for mail return	✓	✓	✓	✓
The percentage of all ACS occupied housing units that receive public assistance income	✓	✓	✓	✓
The percentage of the ACS population aged 5 years and over that speaks a language other than English at home	✓	✓	✓	✓
Proportion non-Hispanic Female	✓	✓	✓	✓
Proportion non-Hispanic American Indians and Native Americans alone	✓	✓	✓	✓
Proportion non-Hispanic White alone	✓	✓	✓	✓
Proportion of occupied units with more than 1.01 persons per room among all occupied units	✓	✓	✓	✓
Proportion of moved households	✓	✓	✓	✓
Proportion of households with one or more people under 18 years old	✓	✓	✓	✓
Proportion of housing units in structures containing 2 to 9 housing units	✓	✓	✓	✓
Interviewer observation				
IO2. Neighborhood type	✓	-	-	-
IO3. Indication of children	✓	✓	-	-
IO4. Estimated number of residents	-	✓	-	-
Virtual observation				
VO8. Does a sidewalk exist in front of the dwelling unit?	-	-	✓	✓
VO9. Is there a driveway or parking lots for the dwelling unit?	✓	-	-	-

[a] In adaptive survey design (ASD) models, information from the sample monitoring paradata (e.g., most recent interim response indicator before final status, number of screener contact attempts before final status) were also added to the models. The weight adjustment models, on the other hand, only included the area-level variables.

with the respondents, and that their best guess on initial observations on the physical appearance of the dwelling and neighborhood was needed. Improvements would include the use of photos in an exercise that asks interviewers to provide their responses based on photos, and then a group review of the responses could be conducted to achieve lower variance across interviewers during the main study. Training can also help in the case of virtual observations, where we encountered challenges similar to those reported by Vercruyssen and Loosveldt (2017), such as unavailability of Street View for some DUs, issues of visual confirmation of the exact house number, out-of-date street views, and challenges with DUs that are in apartment buildings.

5.2 Predictive power of observations

In terms of predictive power, it is interesting to note that four out of the eight interviewer observation items show evidence of significant associations with response propensity. More specifically, four interviewer observation items are associated with the screener response indicator, and two of those same four items are associated with the final interview response indicator. On the other hand, only one of the 16 virtual observation items (VO5 neighborhood type) shows evidence of a significant association with screener response

Table 5
Assessment of value added by interviewer and virtual observations

Model	Original pseudo R^2	Final pseudo R^2	Selected interviewer / virtual observation items
Screener Response Propensity Model			
1. ASD	0.348	0.351	IO2 Neighborhood type; IO3 Indication of children; VO9 Is there a driveway or parking lots for the dwelling unit?
2. Weighting	0.013	0.023	IO3 Indication of children; IO4 Estimated number of residents
Final Interview Response Propensity Model			
3. ASD	0.289	0.299	VO8 Does a sidewalk exist in front of the dwelling unit?
4. Weighting	0.053	0.066	VO8 Does a sidewalk exist in front of the dwelling unit?

Table 6
Number of completed virtual observations and hours worked, by virtual observer

Virtual Observer	# of DUs completed	Time spent
1	839	35 hours
2	845	64 hours
3	850	60 hours
4a	593	34 hours
4b	250	15 hours
Total	3,377	208 hours

The total number of DUs observed by the virtual observers (3,377) is higher than the sampled DUs (2,552) due to overlapping observations of some DUs for the purpose of reliability testing.

propensity. All interviewer observation items are related to at least one of the five survey outcomes; IO5 (house not well kept) is associated with most outcomes. Nine of the 16 virtual observation items are related to some survey outcomes, with VO10 (litter/vandalism) related to most of the outcomes.

Our conclusion is that neighborhood type (either interviewer or virtual) has comparatively higher predictive power (in terms of both response propensity and survey outcomes) than other items. However, there is not much benefit given the other variables that are available. Furthermore, interviewer observations for indication of children and house condition have more predictive power than other interviewer observation items, and all virtual observation items. This could be due to the rich information that interviewers can observe at the site at the time of the survey interview, whereas the virtual observations can only visually obtain limited information with a potential delay in the time frame. Lastly, in the cases of both interviewer and virtual observations, DU type has positive results for completeness, interviewer variance and reliability (virtual observations), and presents some small indications of predictive power; however, there is not enough evidence of validity and association with response indicators. Because of its high quality, DU type should be considered as a candidate for weighting adjustments, with evidence still needed of an association with response indicators in future contexts.

We further assessed the value added to response propensity models under the contexts of ASD and the weighting process. The evaluation shows that there is limited value added by the observation data. For the screener ASD model, interim disposition codes of DUs (e.g., initial refusals) dominate the response propensity model and, therefore, there is limited value added beyond the covariates that exist. The neighborhood type (interviewer), indications of children (interviewer), and existence of a driveway/parking lot (virtual) show the most potential for improving the screener ASD model. For the final interview, whether or not a sidewalk existed (virtual) shows potential benefit for the final interview ASD model. Accessibility to the DU (i.e., existence of a driveway/parking lot, and existence of sidewalk) is not an interviewer observation item but could be easily recorded by interviewers. With the added value from the two items, accessibility items could be added to future interviewer observation forms. It also should be noted that improvement in the quality of these items is needed before consideration for ASD or weighting adjustments. With the screener being a substantial component of total nonresponse and the lack of useful auxiliary information on DUs, DU observation data could potentially reduce nonresponse bias through the weighting process. If quality improves, the screener weighting process could benefit from including the number of residents (interviewer) and indication of children (interviewer), and the final

Table 7
Summary of the evaluation for interviewer observations

				Predictive Power		Value added analysis			
						Screener response propensity model		Final response propensity model	
Item	Completeness[a]	Validity	Interviewer variance	Response indicator[b]	Key survey outcome[b]	ASD	Weighting	ASD	Weighting
Dwelling unit type	Perfect	Fair	Not sig.	0/2	1/5	No	No	No	No
Neighborhood type	Perfect	Fair	Sig.	1/2	2/5	Yes	No	No	No
Indication of children	Perfect (14%)	Fair	Sig.	2/2	1/5	Yes	Yes	No	No
Estimated number of residents	Perfect (20%)	Fair	Sig.	0/2	2/5	No	Yes	No	No
Indication of house not well kept	Perfect	N/A	Sig.	1/2	3/5	No	No	No	No
Indication of house exterior not maintained	Perfect	N/A	Sig.	2/2	2/5	No	No	No	No
Indication of abandoned vehicle	Perfect	N/A	N/A	0/2	1/5	No	No	No	No
Indication of long grass	Perfect	N/A	N/A	0/2	2/5	No	No	No	No

[a] Rating scale for completeness: Perfect: < 1% missing; High: 1–4% missing; Moderate: 4–10% missing; Low: > 10% missing. Numbers in parentheses show percentage of Don't knows. [b] associated/total possible

Table 8
Summary of the evaluation for virtual observations

			Reliability[b]		Predictive Power		Value added analysis			
			Aggreement . . .				Screener response propensity model		Final response propensity model	
Item	Completeness[a]	Validity	among VOs	between VOs and IOs	Response indicator[c]	Key survey outcome[c]	ASD	Weighting	ASD	Weighting
Was Street View available?	High	N/A	Fair	N/A	0/2	0/5	No	No	No	No
Can you visually confirm the dwelling unit number?	High (51%)	N/A	Low	N/A	0/2	0/5	No	No	No	No
Is the dwelling unit visible?	Moderate	N/A	Low	N/A	0/2	1/5	No	No	No	No
What type of dwelling unit?	Moderate	Fair	Fair	Fair	0/2	0/5	No	No	No	No
Is there a locked gate that impedes access to the dwelling unit?	High (15%)	N/A	Low	N/A	0/2	3/5	No	No	No	No
What type of neighborhood?	Moderate	Fair	Fair	Low	1/2	2/5	No	No	No	No
What is your best guess of the number of people living in the dwelling unit?	High (27%)	Low	Low	Low	0/2	3/5	No	No	No	No
Is there any indication that the dwelling unit is not-well-kept?	High (32%)	N/A	Low	Low	0/2	0/5	No	No	No	No
Does a sidewalk exist in front of the dwelling unit?	High (8%)	N/A	Low	N/A	0/2	0/5	No	No	Yes	Yes
Is there a driveway or parking lots for the dwelling unit?	High (8%)	N/A	Low	N/A	0/2	0/5	Yes	No	No	No
Is there evidence of litter or vandalism?	High (29%)	N/A	Low	Low	0/2	4/5	No	No	No	No
How many windows do you see in the front face of the dwelling unit?	High (20%)	N/A	Low	N/A	0/2	2/5	No	No	No	No
Is there any indication of a child or children living in the dwelling unit?	High (37%)	Low	Fair	Low	0/2	2/5	No	No	No	No
Full circle, do you see any non-residential buildings?	High (13%)	N/A	Low	N/A	0/2	3/5	No	No	No	No
How many car lengths is the rooftop width?	High (10%)	N/A	Low	N/A	0/2	3/5	No	No	No	No
Is the street condition good?	High (11%)	N/A	Low	N/A	0/2	0/5	No	No	No	No

[a] Rating scale for completeness: Perfect: < 1% missing; High: 1–4% missing; Moderate: 4–10% missing; Low: > 10% missing. Numbers in parentheses show percentage of Don't knows. [b] Rating scale for agreement rate: Perfect: >99% agreement; High: 95–99% agreement; Moderate 90–95% agreement; Fair: 75–90% agreement; Low: <75% agreement. [c] associated/total possible

interview weight adjustment could potentially benefit from including whether or not a sidewalk existed (virtual).

5.3 Costs of the observations

The focus of this study is on an in-person survey, and it is more efficient to conduct the observations by interviewers; however, if a different mode of data collection is used (e.g., a telephone survey), then virtual observations would be a more cost-efficient relative to sending out someone in-person to make observations. Given the limited value added that we found in this study, we recommend to balance the cost with the value added by the observations, which is affected by the set of covariates that may already be available to help reduce potential nonresponse bias through ASD and/or weighting adjustments. It should be noted that the predictive power of the observation items is likely to be different for other surveys with different survey outcomes, thus a different set of observation items might be more useful. For example, in a crime-related survey, neighborhood characteristics or litter/vandalism might be useful items to retain in the observation forms. Because of the in-person data collection for FoodAPS, there is not much added effort to collect interviewer observations. Therefore it is worth trying to improve upon the quality of the observations and see if additional value can be added to help reduce bias through the adaptive survey design and weighting processes.

5.4 General discussion

Our evaluation provides a specific application to FoodAPS survey outcomes that shows slightly higher predictive power as compared to the Vercruyssen and Loosveldt (2017) virtual observation study, with a similar finding that the pitfalls (i.e., outdated and pixelated images, coverage rate) of the virtual observations might limit the ease of use and the quality of the observations. Interviewer observation items have better completeness, validity and predictive power compared with virtual observation items, and interviewer observations cost less than virtual observations for in-person surveys. Therefore, the interviewer observations have more potential for the FoodAPS in-person survey.

The main limitation of this study is that the quality of the observations in terms of completeness, validity and variance/reliability may affect the predictive power of the observations and attenuate the value added in reducing the nonresponse bias through ASD and weighting processes. In general, for FoodAPS, the application of observation items is now an iterative approach. The first implementation, as we determined from the study, leaves room for improvement on the quality of the estimates. For example, if the interviewer variance of the recorded observations can be reduced, some interviewer observation items may be useful in ASD and/or weighting adjustments for potentially reducing bias due to nonresponse. In particular, we found some predictive power but questionable interviewer variance for neighborhood type, indication of children, household condition, and number of residents. On the other hand, good quality (with exception of validity) but little predictive power is seen for DU type.

A second iteration of the implementation of interviewer observations is needed. As mentioned above, there is still a need to improve the quality of the observations through appropriate training. Possible training methods may include providing relevant cues to interviewers (West & Li, 2019). West and Kreuter (2018) suggested several strategies, such as providing verbal guidance about strategies to avoid and incorporating practice training sessions for recording interviewer observations based on real photographs of housing units and neighborhoods. The appropriate training and useful materials would be a practical and useful way to improve the quality of interviewer observations.

References

Beaumont, J.-F., Bocci, C., & Haziza, D. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, *30*, 607–621.

Brick, J. M., & Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, *33*(3), 735–752. doi:10.1515/JOS-2017-0034

Capanu, M., Gönen, M., & Begg, C. B. (2013). An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Statistics in medicine*, *32*(26), 4550–4566.

Chapman, C. (2014). *National center for education statistics adaptive design overview.* Federal Committee on Statistical Methodology Conference, Washington DC. The Hague: International Statistical Institute.

Clarke, P., Ailshire, J., Melandez, R., Bader, M., & Morenoff, J. (2010). Using google earth to conduct a neighborhood audit: Reliability of a virtual audit instrument. *Health Place*, *16*(6), 1224–1229. doi:10.1016/j.healthplace.2010.08.007

Giangrande, M., Brick, J. M., Morganstein, D., & Lewis, K. (2018). Virtual listing: Gis approaches to improve survey listing efficiency. In *In jsm proceedings, survey research methods section.* Alexandria, VA: American Statistical Association.

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, *72*(2), 167–89.

Hartman, D. P. (1977). Consierations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, *10*(1), 103–116.

Hidiroglou, M., & You, Y. (2016). Comparison of unit leve and area level small area estimators. *Survey Methodology*, *42*(1), 41–61.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, *19*(2), 81–97.

Kirkpatrick, S. I., & Tarasuk, V. (2011). Housing cirumstances are associated with household food access among low-income urban families. *Journal of Urban Health*, *88*(2), 284–296.

Kreuter, F. (2013). *Improving surveys with paradata: Analytic uses of process information.* Hoboken, NJ: John Wiley & Sons, Inc.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., . . . Raghunathan, T. E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for nonresponse: Examples from multiple surveys. *Journal of the Royal Statistical Society Series A—Statistics in Society*, *173*(2), 389–407.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *22*(1), 159–174.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys.* Hoboken, NJ: Wiley.

Little, R. J. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, *54*(2), 139–157.

Little, R. J., & Vartivarian, S. (2005). *Does weighting for nonresponse increase the variance of survey means? survey methodology*.

Lynn, P. (2003). PEDAKSI: Methodology for collection data about survey non-respondents. *Quality & Quantity*, *37*, 239–261.

Mooney, S., Bader, M. D. M., Lovasi, G. S., Teitler, J. O., Koenen, K. C., Aiello, A. E., . . . Rundle, A. G. (2017). Street audits to measure neighborhood disorder: Virtual or in-person? *American Journal of Epidemiology*, *186*(3), 265–273.

Olson, K. (2013). Paradata for nonresponse adjustment. *The Annals of the American Academy of Political and Social Science*, *645*(1), 142–170.

Plewis, I., Calderwood, L., & Mostafa, T. (2017). Can interviewer observations of the interview predict future response? *Methods, data, analysis*, *11*(1), 29–44.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage Publications.

Rosen, J. A., Murphy, J., Peytchev, A., Holder, T., Dever, J. A., Herget, D. R., & Pratt, D. J. (2014). Prioritizing low-propensity sample members in a survey: Implications for nonresponse bias. *Survey Practice*, *7*(1).

SAS Institute Inc. (2013). *Sas 9.4 statements: Reference*. Cary, NC: SAS Institute Inc.

Schräpler, J., Schupp, J., & Wagner, G. G. (2010). Individual and neighborhood determinants of survey nonresponse: An analysis based on a new subsample of the german socio-economic panel, microgeographic characteristics and survey-based interviewer characteristics. SOEP Paper No. 288. Retrieved from http://dx.doi.org/10.2139/ssrn.1588730

Sinibaldi, J., Durrant, G. B., & Kreuter, F. (2013). Evaluating the measurement error of interviewer observed paradata.public opinion quarterly. *77*(S1), 173–193.

Sinibaldi, J., Trappmann, M., & Kreuter, F. (2014). Which is the better investment for nonresponse adjustment: Purchasing commercial auxiliary data or collecting interviewer observations?public opinion quarterly. *78*(2), 440–473.

Smith, T. W. (2011). The report of the international workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys. *International Journal of Public Opinion Research*, *23*(3), 389–402.

Tourangeau, R., Brick, J. M., Lohr, S., & Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(1), 203–223.

U.S. Census Bureau. (2019). Populatin estimates program. Population and housing estimates. Retrieved from https://www.census.gov/housing/hvs/definitions.pdf

Vercruyssen, A., & Loosveldt, G. (2017). Using google street view to validate interviewer observations and predict nonresponse: A belgian case study. *Survey Research Methods*, *11*(3), 345–360.

West, B. T. (2013a). An examination of the quality and utility of interviewer observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society: Series A*, *176*, 211–225.

West, B. T. (2013b). The effects of errors in paradata on weighting class adjustments: A simulation study. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information*. New Jersey: Wiley-Hoboken.

West, B. T., & Kreuter, F. (2013). Factors affecting the accuracy of interviewer observations: Evidence from the national survey of family growth (NSFG). *Public Opinion Quarterly*, *77*(2), 522–548.

West, B. T., & Kreuter, F. (2018). Strategies for increasing the accuracy of interviewer observations of respondent features. *Methodology*, *14*, 16–29.

West, B. T., Kreuter, F., & Trappmann, M. (2014). Is the collection of interviewer observations worthwhile in an economic panel survey? new evidence from the german labor market and social security (PASS) study. *Journal of Survey Statistics and Methodology*, *2*(2), 159–181.

West, B. T., & Li, D. (2019). Sources of variance in the accuracy of interviewer observations. *Sociological Methods & Research*, *48*(3), 485–533.

West, B. T., Wagner, J., Hubbard, F., & Gu, H. (2015). The utility of alternative commercial data sources for survey operations and estimation: Evidence from the national survey of family growth. *Journal of Survey Statistics and Methodology*, *3*(2), 240–264.

Appendix A
Interviewer Variance Model

Hierarchical generalized linear modeling (specifically, multilevel logistic regression) was used to investigate the interviewer variance (Raudenbush & Bryk, 2002), with the interviewer as a level-2 factor and households at level 1. The proposed model for each outcome is written as:

$$\eta_{ij} = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j} + \beta_{1j}X_{1ij} + \ldots + \beta_{Qj}X_{Qij} \quad \text{(Level 1)}$$

$$\beta_{qj} = \gamma_{q0} + u_{qj} \quad \text{(Level 2)},$$

where η_{ij} is the logit for the i^{th} household interviewed by j^{th} interviewer; β_{0j} is the level-1 intercept coefficient, β_{qj} is the level-1 slope coefficient for each level-1 predictor X_{qij}, and X_{qij} is the level-1 predictor; γ_{q0} is the level-2 intercept; and u_{qj} is the random interviewer effect. For each interviewer j, the vector $(u_{0j}, u_{1j}\ldots u_{Qj})'$ is assumed to be multivariate normally distributed, with each random effect u_{qj} having a mean of zero and variance of $\text{Var}\left(u_{qj}\right) = \tau_{qq}$. In this analysis, only the random effects of the intercept and the slopes of the screener questionnaire predictors were allowed to vary. The random effects were assumed independent. The integral approximation method (Laplace approximation) was chosen for the parameter estimation rather than the pseudo-likelihood method, as it provides less-biased estimation (Capanu, Gönen, & Begg, 2013).

Appendix B
Forms

B.1 Observation Checklist—Dwelling Unit (DU) Description

1. DU Type

1. Detached/single family home
2. Modular home
3. Townhouse/rowhouse/duplex/ triplex/quadplex
4. Garden apartment/condo
5. Midrise apartment/condo
6. High rise apartment/condo
7. Detached/single family home converted to apartments
8. Mobile home/trailer/recreational vehicle
9. Student housing – Campus dormitory
10. Student housing – Apartment
11. Hotel/motel room
12. Rooming or boarding house
13. Transitional housing
14. Work camp
15. On base military housing (non-barracks)
16. Assisted living
17. Group home
18. Not a DU (Specify)

2. Neighborhood Type?

1. High income
2. Upper middle income
3. Middle income
4. Low income

3. Is there any indication of a child or children living in the DU? (Please look for evidence indicating the presence of children such as

- baby strollers, outdoor toys/shoes, bikes, swing sets, trampolines, basketball hoop (porch, yard, or driveway)
- car seats, booster seats, or toys in the backseat of cars (driveway)
- baby blankets, toys, furniture, child equipment (open garage or inside the house through window)
- boxes for baby wipes or diapers, candy wrappers, stickers/crayons/miscellaneous kids decorations
- sounds of children)

1. Yes
2. No
3. Don’t know

4. What is your best guess of the number of people living in the DU?

1. 1-2

2. 3+
3. Don't know

5. DU Condition? (choose all that apply)

1. Well kept
2. House exterior not maintained well
3. Abandoned vehicle around
4. Long grass

6. Other Notes:

B.2 Virtual Observation Checklist—Dwelling Unit (DU) Description

Q0 Was Street View available?

1. Yes
2. No

Q1 Can you visually confirm the dwelling unit number?

1. Exact
2. Inferred from other dwelling units
3. Cannot confirm

Q2 Is the dwelling unit visible?

1. Yes
2. No

Q3 What type of dwelling unit?

1. Single/duplex/townhouse
2. Condo/apartments/other

Q4 Is there a locked gate that impedes access to the dwelling unit?

1. Yes
2. No
3. DK

Q5 What type of neighborhood?

1. Low income
2. Middle to high income
   Note: Do not look for data, just observe by looking at the house and street (like an interviewer would)

Q6 What is your best guess of the number of people living in the DU?

1. 1-2
2. 3+
3. DK

Q7 Is there any indication that the dwelling unit is not-well-kept?

1. Yes

2. No
3. DK

Q8 Does a sidewalk exist in front of the dwelling unit?

1. Yes
2. No
3. DK

Q9 Is there a driveway or parking lots for the dwelling unit?

1. Yes
2. No
3. DK

Q10 Is there evidence of litter or vandalism?

1. Yes
2. No
3. DK

Q11 How many windows do you see in the front face of the dwelling unit?

1. ≤ 2
2. > 2
3. DK
4. NA

   Note: Do not count in or around door

Q12 Is there any indication of a child or children living in the DU?

1. Yes
2. No
3. DK

Q13 Full circle, do you see any non-residential buildings?

1. Yes
2. No
3. DK

Q14 How many car lengths is the rooftop width?

1. ≤ 2
2. > 2
3. DK
4. NA

Q15 Is the street condition good?

1. Yes
2. No
3. DK

Appendix C
Tables

Table C1
Frequency tables for the interviewer observation items

Label (coarsened/detailed)	Freq.	%	Cumulative Freq.	Cumulative %
IO1. Interviewer observation				
Condo/apartments/other				
Garden apartment/condo	168	6.80	168	6.80
Midrise apartment/condo	236	9.55	404	16.36
High rise apartment/condo	18	0.73	422	17.09
Detached/single family home converted to apartments	72	2.91	494	20.00
Mobile home/trailer/recreational vehicle	42	1.70	536	21.70
Student housing[a]	24	0.97	560	22.67
Not a dwelling unit	28	1.13	588	23.81
Single/townhouse				
Detached/single family home	1534	62.11	2122	85.91
Modular Home	37	1.50	2159	87.41
Townhouse/rowhouse/duplex/triplex/ quadplex	311	12.59	2470	100.00
IO2. Neighborhood type				
Low income				
Low income	591	23.93	591	23.93
Middle to High income				
High income	151	6.11	742	30.04
Upper middle income	452	18.30	1194	48.34
Middle income	1276	51.66	2470	100.00
IO3. Indication of children				
No	1844	74.66	1844	74.66
Yes	283	11.46	2127	86.11
Don't Know	343	13.89	2470	100.00
IO4. Estimated number of residents				
1–2	1194	48.34	1194	48.34
3+	779	31.54	1973	79.88
Don't Know	497	20.12	2470	100.00
IO5. Indication of house not well kept				
No	2065	83.60	2065	83.60
Yes	405	16.40	2470	100.00
IO6. Indication of house exterior not maintained				
No	2147	86.92	2147	86.92
Yes	323	13.08	2470	100.00
IO7. Indication of abandoned vehicle				
No	2435	98.58	2435	98.58
Yes	35	1.42	2470	100.00
IO8. Indication of long grass				
No	2422	98.06	2422	98.06
Yes	48	1.94	2470	100.00

[a] Apartment, hotel/motel room, rooming or boarding house, transitional housing, assisted living

Table C2
Estimated associations of interviewer observations with related survey items and ACS data

Survey item/ACS variables	Interviewer Observation	Sample size	Kappa Coefficient	Odds Ratio	Specificity	Sensitivity
Proportion of population with single-family house in SSU—ACS (High vs. Low)	IO1. Dwelling unit type (Single vs. Condo)	2470	0.31	7.44*	0.92	0.38
Reported household income in screener—Survey (High vs. Low)	IO2. Neighborhood type (Middle/High vs. Low)	771	0.39	10.46*	0.93	0.42
Proportion of population above 185% poverty line in SSU—ACS (High vs. Low)		2470	0.41	8.32*	0.88	0.53
Household having children under age 5—Survey (Yes vs. No)	IO3. Indication of children (Yes vs. No)	647	0.20	3.22*	0.32	0.87
Household having children under age 18—Survey (Yes vs. No)		405	0.23	4.30*	0.32	0.90
Household size—Survey (3+ vs. 1–2)	IO4. Estimated number of residents (3+ vs. 1–2)	682	0.21	2.39*	0.51	0.69

* $p < 0.05$

Table C3
Hierarchical generalized linear models: specification and variance component estimates

		Level 2 Components Variance Estimates	
Outcome	Level 1 Covariates	Coef.[a]	S.E.
Dwelling unit type	Constant	0.27	0.33
	Reported household income in screener	0.57*	0.41
	Prop. of population above 185% poverty line in SSU	-	-
	Prop. of population with single-family house in SSU	-	-
Neighborhood type	Constant	1.78*	0.71
	Reported household income in screener	0.38	0.41
Indication of children (Model A)	Constant	0.46*	0.33
	HH with children under age 18 in survey	not estimable	
	Prop. of population above 185% poverty line in SSU	-	-
	Prop. of population under 18 in SSU	-	-
Indication of children (Model B)	Constant	0.43	0.25
	HH with children under age 5 in survey	not estimable	
	Prop. of population under 18 in SSU	-	-
Estimated number of residents	Constant	0.86*	0.30
	Household size in survey	not estimable	
	Household size in SSU	-	-
Indication of house not well kept	Constant	0.34*	0.30
	Reported household income in screener	0.23	0.36
	Prop. of population above 185% poverty line in SSU	-	-
Indication of house exterior not maintained	Constant	0.55*	0.26
	Reported household income in screener	not estimable	
	Prop. of population above 185% poverty line in SSU	-	-

[a] Only selected level-1 covariates are allowed to vary. Some estimates may not be estimable due to some cells in the cross-tabulation being empty or only containing one observation.
* $p < 0.05$

Table C4
Logistic regression coefficients from the response propensity model, with interviewer observations as predictors

		Odds Ratio	
Interviewer oberservation	Sample size	Screener response indicator	Final interview response indicator
IO1. Dwelling unit type (Single vs. Condo)	2143	0.96	0.98
IO2. Neighborhood type (Middle/High vs. Low)	2143	0.76*	0.78
IO3. Indication of children (Yes vs. No)	1845	1.64*	1.70*
IO4. Estimated number of residents (3+ vs. 1–2)	1742	1.01	1.12
IO5. Indication of house not well kept (Yes vs. No)	2143	1.46*	1.25
IO6. Indication of house exterior not maintained (Yes vs. No)	2143	1.52*	1.39*
IO7. Indication of abandoned vehicle (Yes vs. No)	2143	1.64	1.08
IO8. Indication of long grass (Yes vs. No)	2143	1.10	1.17

* $p < 0.05$

Table C5
Logistic regression coefficients from the model of food adequacy, with interviewer observations as predictors

Interviewer observation	Sample size	Odds Ratio
IO1. Dwelling unit type (Single vs. Condo)	421	2.60*
IO2. Neighborhood type (Middle/High vs. Low)	421	4.89*
IO3. Indication of children (Yes vs. No)	369	0.79
IO4. Estimated number of residents (3+ vs. 1–2)	421	3.14
IO5. Indication of house not well kept (Yes vs. No)	421	0.27
IO6. Indication of house exterior not maintained (Yes vs. No)	421	0.21*
IO7. Indication of abandoned vehicle (Yes vs. No)	421	0.10
IO8. Indication of long grass (Yes vs. No)	421	0.53

* $p < 0.05$

Table C6
Linear regression coefficients and adjusted R-squares from models for various survey items, with the interviewer observations as predictors

	Number of FAFH items		Number of FAFH items		Number of FAFH items		Number of FAFH items	
Interviewer Observation	Coeff.	Adj. R^2	Coeff.	Adj. R^2	Coeff.	Adj. R^2	Coeff.	Adj. R^2
IO1. Dwelling unit type (Single vs. Condo)	−1.08	0.002	−1.45	0.002	−6.07	0.008	−2.79	0.001
IO2. Neighborhood type (Middle/High vs. Low)	−1.46	0.005	−1.24	0.001	−11.54*	0.034	−1.63	−0.001
IO3. Indication of children (Yes vs. No)	0.44	−0.002	−2.58*	0.009	1.10	−0.003	−7.50	0.019
IO4. Estimated number of residents (3+ vs. 1–2)	−0.07	−0.003	−2.46*	0.016	−0.12	−0.003	−4.24*	0.011
IO5. Indication of house not well kept (Yes vs. No)	1.31	0.002	2.41*	0.007	11.41*	0.022	4.75*	0.005
IO6. Indication of house exterior not maintained (Yes vs. No)	1.39	0.002	1.68	0.002	11.20*	0.019	3.42	0.001
IO7. Indication of abandoned vehicle (Yes vs. No)	1.84	−0.002	−10.72	0.010	18.64*	0.002	−38.98	0.032
IO8. Indication of long grass (Yes vs. No)	1.21	−0.002	3.53*	<.001	6.13	−0.002	8.22*	<.001

A simple weighted regression was run for each outcome and interviewer observation combination.
All interviewer observation variables are dichotomized in this table.
* $p < 0.05$

Table C7
Frequency distributions of the virtual observation items

			Cumulative	
	Freq.	%	Freq.	%
VO0. Was Street View available?				
No	466	18.87	466	18.87
Yes	1921	77.8	2387	96.68
Missing	82	3.32	2469	100.00
VO1. Can you visually confirm the dwelling unit number?				
Exact	937	37.95	937	37.95
Inferred from other dwelling units	199	8.06	1136	46.01
Cannot confirm	1249	50.59	2385	96.6
Missing	84	3.4	2469	100.00
VO2. Is the dwelling unit visible?				
No	559	22.64	559	22.64
Yes	1751	70.92	2310	93.56
Missing	159	6.44	2469	100.00
VO3. What type of dwelling unit?				
Condo/apartments/other	585	23.69	585	23.69
Single/duplex/townhouse	1784	72.26	2369	95.95
Missing	100	4.05	2469	100.00
VO4. Is there a locked gate that impedes access to the dwelling unit?				
No	1893	76.67	1893	76.67
Yes	135	5.47	2028	82.14
DK	358	14.5	2386	96.64
Missing	83	3.36	2469	100.00
VO5. What type of neighborhood?				
Low income	281	11.38	281	11.38
Middle to high income	2016	81.65	2297	93.03
Missing	172	6.96	2469	100.00
VO6. What is your best guess of the number of people living in the dwelling unit?				
1–2	729	29.53	729	29.53
3+	1001	40.54	1730	70.07
DK	656	26.57	2386	96.64
Missing	83	3.36	2469	100.00
VO7. Is there any indication that the dwelling unit is not-well-kept?				
No	1527	61.85	1527	61.85
Yes	61	2.47	1588	64.32
DK	797	32.28	2385	96.6
Missing	84	3.4	2469	100.00
VO8. Does a sidewalk exist in front of the dwelling unit?				
No	718	29.08	718	29.08
Yes	1479	59.9	2197	88.98
DK	189	7.65	2386	96.63
Missing	83	3.36	2469	100.00
VO9. Is there a driveway or parking lots for the dwelling unit?				
No	87	3.52	2275	3.52
Yes	2105	85.26	2188	88.78
DK	194	7.86	2469	96.64
Missing	83	3.36	2469	100.00

Continues on next page

Continued from previous page

			Cumulative	
	Freq.	%	Freq.	%
VO10. Is there evidence of litter or vandalism?				
No	1612	65.29	1745	65.29
Yes	48	1.94	133	67.23
DK	724	29.32	2469	96.55
Missing	85	3.44	2469	100.00
VO11. How many windows do you see in the front face of the dwelling unit?				
<= 2	497	20.13	580	20.13
> 2	996	40.34	1576	60.47
DK	493	19.97	2069	80.44
NA	400	16.2	2469	96.64
Missing	83	3.36	2469	100.00
VO12. Is there any indication of a child or children living in the dwelling unit?				
No	1319	53.42	1558	53.42
Yes	156	6.32	239	59.74
DK	911	36.9	2469	96.64
Missing	83	3.36	2469	100.00
VO13. Full circle, do you see any non-residential buildings?				
No	1771	71.73	2159	71.73
Yes	305	12.35	388	84.08
DK	310	12.56	2469	96.64
Missing	83	3.36	2469	100.00
VO14. How many car lengths is the rooftop width?				
<= 2	818	33.13	901	33.13
> 2	947	38.36	1848	71.49
DK	238	9.64	2086	81.13
NA	383	15.51	2469	96.64
Missing	83	3.36	2469	100.00
VO15. Is the street condition good?				
No	50	2.03	2200	2.03
Yes	2066	83.68	2150	85.71
DK	269	10.9	2469	96.61
Missing	84	3.4	2469	100.00

There is one DU with completed interviewer observation but no completed virtual observation, resulting in the total completed virtual observation case to be 2,469.

Table C8
Estimated associations of virtual observations with related survey items and ACS data

Survey Item/ACS Variable	Virtual Observation		N	κ	O.R.	Specifity	Sensitivity
	Name	Categ.					
Proportion of population with single-family house in SSU—ACS (High vs. Low)	VO3. Dwelling unit type	Single/townhouse vs. Condo	2469	0.38	12.27*	0.94	0.43
		Missing vs. Condo	-	-	11.81*	-	-
Reported household income in screener—Survey (High vs. Low)	VO5. Neighborhood type	Middle/High vs. Low income	771	0.22	7.14*	0.96	0.24
		Missing vs. Low income	-	-	3.40*	-	-
Proportion of population above 185% poverty line in SSU—ACS (High vs. Low)	VO5. Neighborhood type	Middle/High vs. Low income	2469	0.28	6.59*	0.94	0.29
		Missing vs. Low income	-	-	5.56*	-	-
Household having children under age 5—Survey (Yes vs. No)	VO12. Indication of children	Yes vs. No	726	< 0.01	1.01	0.14	0.86
		Missing/DK vs. No	-	-	0.84	-	-
Household having children under age 18—Survey (Yes vs. No)	VO12. Indication of children	Yes vs. No	463	0.14	3.89*	0.20	0.94
		Missing/DK vs. No	-	-	0.71	-	-
Household size—Survey (3+ vs. 1–2)	VO6. Estimated number of residents	3+ vs. 1–2	815	0.09	1.47*	0.66	0.43
		Missing/DK vs. 1–2	-	-	0.86	-	-

Agreement rate, Kappa coefficient, specificity and sensitivity are only restricted to the comparisons between the non-missing/DK categories.
* $p < 0.05$

Table C9
Agreement of virtual observation items among virtual observers

Virtual Observation	Agreement rate (%)
VO0. Was Street View available?	87.17
VO1. Can you visually confirm the dwelling unit number?	74.21
VO2. Is the dwelling unit visible?	68.93
VO3. What type of dwelling unit?	86.16
VO4. Is there a locked gate that impedes access to the dwelling unit?	66.42
VO5. What type of neighborhood?	76.10
VO6. What is your best guess of the number of people living in the dwelling unit?	54.21
VO7. Is there any indication that the dwelling unit is not-well-kept?	72.20
VO8. Does a sidewalk exist in front of the dwelling unit?	72.08
VO9. Is there a driveway or parking lots for the dwelling unit?	74.84
VO10. Is there evidence of litter or vandalism?	70.06
VO11. How many windows do you see in the front face of the dwelling unit?	55.97
VO12. Is there any indication of a child or children living in the dwelling unit?	76.10
VO13. Full circle, do you see any non-residential buildings?	68.18
VO14. How many car lengths is the rooftop width?	48.81
VO15. Is the street condition good?	69.69

Table C10
Agreement among virtual observers and interviewers

Virtual Observation	Agreement rate (%)
Dwelling unit type (IO1 & VO3)	83.96
Neighborhood type (IO2 & VO5)	72.58
Indication of children (IO3 & VO12)	55.53
Estimated number of residents (IO4 & VO6)	39.45
Indication of house not well kept (IO5 & VO7)	54.68
Indication of house exterior not maintained (IO6 & VO7)	56.30
Indication of abandoned vehicle (IO7 & VO10)	64.93
Indication of long grass (IO8 & VO10)	64.44

Table C11
Logistic regression odds ratios from the response propensity model, with virtual observations as predictors

Virtual Observation	Response indicator for Screener Odds Ratio	Response indicator for final interview Odds Ratio
VO0. Was Street View available?		
Yes vs. No	0.96	1.22
Missing vs. No	1.33	1.84
VO1. Can you visually confirm the dwelling unit number?		
Inferred vs. Exact	1.24	1.05
Missing/Cannot confirm vs. Exact	1.00	0.87
VO2. Is the dwelling unit visible?		
Yes vs. No	0.98	1.02
Missing vs. No	1.18	1.50
VO3. What type of dwelling unit?		
Single vs. Condo	0.97	0.97
Missing vs. Condo	1.16	1.30
VO4. Is there a locked gate that impedes access to the dwelling unit?		
Yes vs. No	0.98	0.88
Missing/DK vs. No	1.06	1.06
VO5. What type of neighborhood?		
Middle/High vs. Low income	0.68*	0.81
Missing vs. Low income	1.08	0.99
VO6. What is your best guess of the number of people living in the dwelling unit?		
3+ vs. 1–2	1.06	0.95
Missing/DK vs. 1–2	1.05	0.94
VO7. Is there any indication that the dwelling unit is not-well-kept?		
Yes vs. No	1.63	0.81
Missing/DK vs. No	1.05	0.94
VO8. Does a sidewalk exist in front of the dwelling unit?		
Yes vs. No	1.00	1.28
Missing/DK vs. No	1.12	1.51
VO9. Is there a driveway or parking lots for the dwelling unit?		
Yes vs. No	0.86	1.02
Missing/DK vs. No	0.80	1.08
VO10. Is there evidence of litter or vandalism?		
Yes vs. No	0.93	1.29
Missing/DK vs. No	0.94	0.87
VO11. How many windows do you see in the front face of the dwelling unit?		
> 2 vs. <= 2	0.92	0.90
Missing/DK/NA vs. <= 2	0.89	0.82
VO12. Is there any indication of a child or children living in the dwelling unit?		
Yes vs. No	1.37	1.29
Missing/DK vs. No	1.01	0.89
VO13. Full circle, do you see any non-residential buildings?		
Yes vs. No	1.12	1.11
Missing/DK vs. No	0.93	0.96
VO14. How many car lengths is the rooftop width?		
> 2 vs. <= 2	0.91	0.82
Missing/DK /NA vs. <= 2	1.00	0.87
VO15. Is the street condition good?		
Yes vs. No	1.04	0.75
Missing/DK vs. No	1.29	0.89

* $p < 0.05$

Table C12
Logistic regression odds ratios from the model of food adequacy, with virtual observations as predictors

Virtual Observation	Sample size	Odds Ratio
VO0. Was Street View available?		
Yes vs. No	403	1.00
Missing vs. No		0.38
VO1. Can you visually confirm the dwelling unit number?		
Inferred vs. Exact	421	0.54
Missing/Cannot confirm vs. Exact		0.58
VO2. Is the dwelling unit visible?		
Yes vs. No	421	3.20*
Missing vs. No		1.51
VO3. What type of dwelling unit?		
Single vs. Condo	421	2.45
Missing vs. Condo		0.73
VO4. Is there a locked gate that impedes access to the dwelling unit?		
Yes vs. No	421	0.17
Missing/DK vs. No		0.54
VO5. What type of neighborhood?		
Middle/High vs. Low income	421	2.39
Missing vs. Low income		0.69
VO6. What is your best guess of the number of people living in the dwelling unit?		
3+ vs. 1–2	421	1.89
Missing/DK vs. 1–2		0.25*
VO7. Is there any indication that the dwelling unit is not-well-kept?		
Yes/Missing/DK vs. No	421	0.46
VO8. Does a sidewalk exist in front of the dwelling unit?		
Yes vs. No	421	0.67
Missing/DK vs. No		0.73
VO9. Is there a driveway or parking lots for the dwelling unit?		
Yes vs. No	421	2.83
Missing/DK vs. No		2.16
VO10. Is there evidence of litter or vandalism?		
Yes vs. No	421	2.36
Missing/DK vs. No		0.43
VO11. How many windows do you see in the front face of the dwelling unit?		
> 2 vs. <= 2	421	2.65
Missing/DK/NA vs. <= 2		0.60
VO12. Is there any indication of a child or children living in the dwelling unit?		
Yes vs. No	421	0.69
Missing/DK vs. No		0.22*
VO13. Full circle, do you see any non-residential buildings?		
Yes vs. No	421	0.38
Missing/DK vs. No		0.53
VO14. How many car lengths is the rooftop width?		
> 2 vs. <= 2	421	3.01
Missing/DK /NA vs. <= 2		0.59
VO15. Is the street condition good?		
Yes vs. No	421	1.25
Missing/DK vs. No		0.82

* $p < 0.05$

Table C13
Estimated coefficients and R-squared from models regressing number of events/items on virtual observations

	Number of FAH items		Number of FAFH items		Number of FAFH items		Number of FAFH items	
Virtual Observation	Coeff.	Adj. R^2	Coeff.	Adj. R^2	Coeff.	Adj. R^2	Coeff.	Adj. R^2
VO0. Was Street View available?								
Yes vs. No	−2.70	0.016	−1.37	0.003	2.40	−0.003	−1.70	> −.001
Missing vs. No	−2.89		−4.25		4.31		−7.48	
VO1. Can you visually confirm the dwelling unit number?								
Inferred vs. Exact	−1.11	<.001	−0.85	−0.004	−3.86	> −.001	0.82	−0.005
Missing/Cannot confirm vs. Exact	0.74		−0.04		−3.41		−0.10	
VO2. Is the dwelling unit visible?								
Yes vs. No	−1.69	0.005	0.58	−0.003	2.75	−0.002	2.09	−0.001
Missing vs. No	−1.61		−0.62		5.17		−1.15	
VO3. What type of dwelling unit?								
Single vs. Condo	0.42	−0.004	1.00	0.003	6.64	0.009	2.28	0.0
Missing vs. Condo	−0.44		−2.56		6.14		−4.78	
VO4. Is there a locked gate that impedes access to the dwelling unit?								
Yes vs. No	−2.79*	0.014	−3.13*	0.003	−8.22	0.002	−6.40*	0.002
Missing/DK vs. No	1.54		> −0.001		0.74		0.05	
VO5. What type of neighborhood?								
Middle/High vs. Low income	2.87*	0.012	2.78	0.007	13.60*	0.026	3.91	<.001
Missing vs. Low income	3.68*		0.85		18.36*		0.34	
VO6. What is your best guess of the number of people living in the dwelling unit?								
3+ vs. 1–2	2.06*	0.009	1.68	0.003	12.18*	0.037	3.38	0.005
Missing/DK vs. 1–2	1.27		0.20		4.00		−0.67	
VO7. Is there any indication that the dwelling unit is not-well-kept?								
Yes vs. No	−0.48	−0.001	−0.18	−0.005	−4.17	−0.001	−1.71	−0.004
Missing/DK vs. No	0.95		0.07		−3.01		−1.07	
VO8. Does a sidewalk exist in front of the dwelling unit?								
Yes vs. No	−2.33	0.014	−1.82	0.003	−6.03	0.005	−3.97	0.004
Missing/DK vs. No	−1.99		−1.99		−3.51		−4.70	
VO9. Is there a driveway or parking lots for the dwelling unit?								
Yes vs. No	0.79	−0.004	−2.23	0.001	6.60	−0.002	−5.13	0.001
Missing/DK vs. No	0.93		−4.12		9.35		−9.21	
VO10. Is there evidence of litter or vandalism?								
Yes vs. No	−3.26*	0.007	−4.46*	<.001	−15.94*	0.003	−9.45*	0.002
Missing/DK vs. No	1.28		−0.42		−0.84		−2.17	
VO11. How many windows do you see in the front face of the dwelling unit?								
> 2 vs. <= 2	0.88	0.004	1.91*	0.002	2.52	−0.003	4.40*	0.003
Missing/DK/NA vs. <= 2	1.76		1.48		0.25		1.90	
VO12. Is there any indication of a child or children living in the dwelling unit?								
Yes vs. No	0.64	0.002	6.05*	0.025	5.54	0.001	10.42	0.016
Missing/DK vs. No	1.21		0.48		−2.18		−0.65	
VO13. Full circle, do you see any non-residential buildings?								
Yes vs. No	−1.49*	<.001	−2.64*	0.012	−4.15	> −.001	−5.93*	0.015
Missing/DK vs. No	−0.23		−2.39*		−2.76		−5.96*	
VO14. How many car lengths is the rooftop width?								
> 2 vs. <= 2	1.99*	0.010	2.60*	0.012	11.29*	0.032	4.58	0.009
Missing/DK /NA vs. <= 2	1.49		0.75		3.60		−0.21	
VO15. Is the street condition good?								
Yes vs. No	1.69	−0.003	1.35	−0.0004	1.43	−0.005	2.97	<.001
Missing/DK vs. No	2.21		−0.20		2.43		−0.60	

* $p < 0.05$