Assessing Interviewer Observations in the NHIS

Rachael Walsh[1], James Dahlhamer[2], and Nancy Bates[1]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233
[2]National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782

Abstract

Face-to-face surveys provide interviewers the opportunity to gather information beyond the scope of the survey. These observational paradata may prove useful for streamlining the data collection process and enhancing post collection weighting adjustments, both improving the quality of survey estimates. However, interviewer observations can be error-prone especially when the observation requires a judgment based on limited information. This research uses survey paradata to evaluate new interviewer observations made during the collection of the 2013 National Health Interview Survey (NHIS). Multilevel models discern the variation in compliance with the new task attributable to interviewers, finding that the characteristics of the case, not the interviewer, reduced variation in compliance. We then examine variation in the observations by whether contact was made when recording the observations, finding inferential but not factual observations vary based on making contact during the observation. When interacted with contact, 11 of the 15 observations are predictive of either the number of contact attempts made on the case or the odds of refusal—two measures of level of effort.

Key Words: paradata; interviewer observations; NHIS

1. Introduction

Interviewer observations of both responding and nonresponding sample units are a potentially useful form of paradata—that is data about the data collection process—when they correlate with both the propensity to respond and the key survey estimates (Kreuter et al, 2010; West, 2013). When meeting both these criteria, this form of paradata has potential for use in nonresponse weighting adjustments (Bethlehem, 2002; Groves, 2006; Little and Vartivarian, 2005; Peytchev and Olson, 2007). These same criteria apply when evaluating paradata for use in responsive or adaptive design models, which may use auxiliary data in daily response propensity models for case prioritization (Axinn, Link, and Groves, 2011; Groves and Heeringa, 2006; Mohl and LaFlamme, 2007). In the context of case prioritization, adaptive design models balance responses by different key subgroups or domains related to the survey variables of interest.

The growing body of literature suggests observations recorded by interviewers often meet one of these two criteria but not both, and are subject to variation resulting in measurement error (Blom et al., 2011; Durrant et al., 2012; Groves and Heeringa, 2006; Kreuter et al., 2010; Casas-Cordero et al., 2013). Unfortunately, observations recorded by interviewers are often error-prone, specifically when the observations require interviewers to make inferences based on limited information (West, 2013; West and Kreuter, 2013). Observations, specifically inferentially based observations vary in both completion and quality. To further the exploration of these measures, the National Health Interview Survey (NHIS) piloted a set of interviewer observations. As a

Disclaimer: The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau or the National Center for Health Statistics.

preliminary step in assessing the utility of these observations, we address the following research questions:

- Is there an interviewer effect on complying with the new task of collecting observations during contact attempts?
- Are there variations in the interviewer observations by whether they were recorded during a personal visit that resulted in contact versus a noncontact?
- Do interviewer observations predict the level of effort the interviewer puts forth to complete an interview?

2. Data and Methods

2.1 Sampling

The National Health Interview Survey (NHIS) is a cross-sectional in-person household health survey conducted by the U.S. Census Bureau for the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). The target population for the NHIS is the civilian noninstitutionalized population of the United States. Fielded virtually continuously since 1957, the survey produces nationally representative data on health insurance coverage, health care access and utilization, health status, health behaviors, and other health-related topics. Data are collected by roughly 750 trained interviewers with the U. S. Census Bureau using computer assisted personal interviewing (CAPI).

The NHIS questionnaire consists of a core set of questions that remain relatively unchanged from year to year, and supplemental questions that vary annually to collect additional data pertaining to current health issues of national importance. The core instrument has four main modules: Household Composition, Family, Sample Child, and Sample Adult. For the household composition module, a household respondent provides basic sociodemographic information on all members of the household. Within each family, a family respondent who provides health information on each member of the family completes the family module. Additional health information is subsequently collected from the parent or guardian of one randomly selected child under age 18 (the "sample child") and from one randomly selected adult (the "sample adult") aged 18 years or older.

2.2 Paradata

The NHIS collects and utilizes paradata for various aspects of its quality assurance/control program, including assessments of data quality (e.g., nonresponse bias analyses) and monitoring of interviewer performance, with a future goal of applying paradata to adaptive or responsive survey design. The bulk of NHIS paradata to date has been collected with the Contact History Instrument (CHI). The CHI is a fully automated survey instrument developed by the U. S. Census Bureau and first fielded with the 2004 NHIS.[1] Census Bureau interviewers use CHI to record information on each contact attempt made on a household, regardless of outcome.

[1] The survey instrument is programmed using Blaise software. For more information refer to http://www.blaise.com/

In addition to basic information such as date and time and mode of attempt (in-person, telephone), interviewers report the outcome of the attempt (contact with a sample unit member, contact with a non-sample unit member, noncontact) and strategies employed before, during, or immediately after the attempt (e.g., left an appointment card, checked with neighbors, left promotional packet). For attempts resulting in a contact, interviewers complete a screen with 23 categories of verbal and nonverbal concerns and behaviors that may be expressed during interviewer-respondent interactions. Examples include "privacy concerns," "anti-government concerns," "too busy," and "hangs ups-slams door." Other screens collect information on the reasons for a noncontact and why a contact did not result in an interview (e.g., inconvenient time, respondent is reluctant, language barrier).

While useful for characterizing interviewer effort and understanding the types and extent of household reluctance, the general conclusion for the NHIS is that CHI-based measures fall short of meeting the criteria for reduction and correction of nonresponse bias in key health outcomes. That is, while many of the CHI variables are strong predictors of response, these same predictors are weakly correlated, at best, with key NHIS health outcomes (Dahlhamer, 2012; Maitland et al., 2008). This limits the data's utility for guiding on-going data collection decisions (e.g., balancing response among subgroups who differ on key survey outcomes) and making post-collection weight adjustments.

Understanding these limitations, survey methodologists from the Census Bureau and survey sponsoring agencies developed 15 interviewer observation measures that may correlate with both survey response and critical survey-specific estimates. (For a thorough overview of the development process, see Miller et al., 2013.) For example, a question on the presence of a wheelchair ramp or other indications that sample unit residents may have functional limitations or be disabled is hypothesized to impact response to and key health outcomes on the NHIS. Table 1 presents a description of the 15 observation measures, hypotheses as to whether the observations will predict response propensity and specific survey estimates, and the survey that generates each critical estimate.

Only two observations are not expected to correlate with response propensity directly—presence of an adult bicycle and evidence of smoking. Indirectly, however, interviewers use this as an indication that the sample unit is occupied. Also, only one observation is not expected to correlate with any of the key survey estimates—the presence of an access barrier, denoted buzzer/keycode/doorman. Access barriers strongly correlate with response propensity and the CHI currently only records these in very specific situations.

The observation questions were programmed in a stand-alone survey instrument that is invoked when interviewers record contact histories for the *first personal visit* attempt. Interviewers are instructed to record the observations on the first personal visit on which they can observe the sample unit or building within which the sample unit resides. Observations are recorded first (and only once), followed by the usual contact history items captured in CHI. Once the observations are recorded for the sample unit, the questions no longer appear. The goal is to ensure that the observations are collected in a timely fashion, to potentially facilitate daily management of data collection, and comparable across ALL cases, responding and nonresponding, to ensure any observed associations with response and key survey outcomes are not simply artifacts of the measurement process. NHIS interviewers received both a self-study and classroom training on the new observation measures in November and December of 2012, respectively (see Miller et al., 2013, for more information).

Table 1. Interviewer Observations Tested in 2013 NHIS

Observation	Correlated with Response?	Survey Estimate Correlation?	Survey(s)
Graffiti	Yes	Crime victimizations	NCVS
Condition of Sample Unit	Yes	Housing unit condition	AHS
Buzzer/keycode/doorman	Yes	NO	AHS
Well tended yard/garden	Yes	Home value, condition	AHS
Peeling paint/damaged walls	Yes	Home value, condition	AHS
Bars on windows	Yes	Crime victimizations	NCVS
Multiple door locks	Yes	Crime victimizations	NCVS
Presence of children <6	Yes	Flu shot	NHIS
Wheelchair ramp/disabled	Yes	Health status	NHIS
Adult bicycle	No	Health status, crime	NHIS, NCVS
Evidence of smokers	No	Health status	NHIS
Household Income bracket	Yes	Program participation Health insurance	SIPP
Employed adult(s)	Yes	Health insurance, employment	SIPP, CPS
Language other than English	Yes	Other language at home Health status, Medicaid	ACS
Household aged 65+	Yes	Employment	NHIS, CPS

2.3 Analysis Plan

Utilizing NHIS paradata collected from January through April 2013, this research begins with an analysis of interviewer compliance with this new task, moves into an initial assessment of the quality of performing the task, and then ends with the utility of the paradata collected. We ran several different models on the observation records obtained from a sample that included attempted interviews with 26,276 sample units. Restarting a case and case reassignment to a different interviewer resets the observation completion flag, resulting in multiple interviewer observation records for some cases. Our sample included 21,046 cases with only 1 observation record and 3,802 cases with two to five observation records. The 3,802 cases with multiple interviewer observation records generated 8,155 records for 29,201 interviewer observation records. With 1,428 cases without any interviewer observation records, the overall sample size for the analysis is 30,629 records.

2.3.1 Compliance

We looked at two measures of interviewer compliance with the observation protocol—completing the observations and completing the observations on the first personal visit. The results of these measures indicate high levels of compliance. Overall, 95 percent of the cases contained at least one interviewer observation record, with 93 percent of those records recorded on the first personal visit. Interviewer compliance is a key measure of interest given the likelihood of increased error terms resulting from missing data (Lynn, 2003). The two compliance metrics—completing the observation and recording the observation on the first personal visit—were coded as dichotomous indicators at the observation record level. We ran Multilevel logistic regression models, assessing the variation in compliance that is attributable to interviewers themselves (Rabe-Hesketh and Skrondal, 2005; Raudenbush and Bryk, 2002). To explain some of the variation across interviewers, we included interviewer and case characteristics in the models (West and Kreuter, 2013).

Because we are interested in using these observations for nonresponse adjustments, interviewer training instructed interviewers to record the observations on the first personal visit attempt, regardless of contact. To use these observations in adaptive design modeling, they must be available early in the interview period. Both the completion on any contact attempt and completion during the first personal visit attempt were examined in relation to unit accessibility. This included situations where interviewers were denied access to the sample unit, the unit was found to be vacant or unoccupied, or residents refused to participate, all of which are represented in the models. The models also included factors affecting the ability to record observations that are beyond the interviewer's control, such as situations where interviewers cannot locate the sample unit, impassable roads, etc.[2]

Though previous research findings are inconsistent with respect to the role of interviewer characteristics (Sinibaldi, et al, 2013; West and Kreuter, 2013), the models included seven interviewer characteristics—education, gender, experience, supervisory status, position, caseload size, and working additional surveys. The measures of interviewing experience included the years of interviewing both for the Census Bureau and specifically with the NHIS. The 858 interviewers included in the sample had a range of experience, from none to over 30 years. The caseload measure in the models is the number of cases completed from January through the end of April 2013. The majority of interviewers had a high school diploma or equivalent (61 percent) and was female (73 percent). While supervisory interviewers can conduct interviews, this was typically not the case (16 percent), and most interviewers worked more than one survey while fielding the NHIS (73 percent).

2.3.2 Variation by Contact Status

Future analyses will use the NHIS survey response data to assess the accuracy of the observations as well as the relationship between interviewer observations and key survey variables.[3] This analysis will use paradata to assess the variation in the observations collected, as well as the effect of making contact when recording the observations on the ability of the observations to predict level of effort. For the observations to be useful for either nonresponse adjustments or adaptive design, they should be comparable for responding and nonresponding units (Lessler and Kalsbeck, 1992). We assessed the observations based on making contact with a sample unit member versus noncontact.[4] Additionally, we assessed the observations based on concerns expressed by interviewers during centralized training and debriefing sessions.

The final portion of this analysis regressed two measures of level of effort—the number of contact attempts and potential refusal conversions—on the interviewer and case characteristics, as well as the interviewer observations. As these observations could benefit adaptive design models, these measures of level of effort directly relate to the

[2] These are referred to as Other Eligible Noninterview situations.

[3] NHIS is still in the field and therefore the survey response data are not yet available.

[4] Making contact with a nonsample unit member is considered a noncontact.

utility of these observations (Biemer et al, 2012). Number of contact attempts is a count variable. Using the likelihood ratio test, we determined the measure did not have a Poisson distribution, and over dispersion was an issue. Negative binomial regression with the log-link function was the most appropriate regression model for these data (Agresti and Finlay, 2009).

In the CHI, interviewers can record respondent reluctance as a reason they were unable to conduct an interview. For the second model assessing level of effort, we dichotomously recoded cases such that an indication of reluctance during any contact attempt resulted in identifying the case as an interim refusal. As an additional measure of the dependence of the observations on making contact, we stratified by contact during the observation attempt. If direction or magnitude changed, we included an interaction term for the respective level of effort model.

Ideally, the results of this portion of the analysis would show that the observations are independent of making contact during the attempt when recorded (Miller et al., 2013). Additionally, *a priori* contact history research suggests an expected relationship between the level of effort expended and some of the observations tested. For example, cases with access barriers should require more contact attempts than those without barriers, while sample units with wheelchair ramps and all members over the age of 65 should require fewer contact attempts (Atrostic et al., 2001; Bates et al., 2008). Overall, these analyses were an initial assessment of the interviewer observations, looking at compliance with the new task then assessing the quality and utility of the measures.

3. Findings

3.1 Interviewer Compliance

Table 2 displays the odds ratios of the multilevel model predicting compliance at the observation record level. The attributes of the sample unit itself were the primary drivers of compliance with recording the observations, which is consistent with previous research (West and Kreuter, 2013). The intraclass correlation coefficient (ICC) shows the variation attributable to interviewers. For the model predicting presence of the observations, 33 percent of the variation in recording observations was attributable to variation across interviewers. Taking into consideration the interviewer and case characteristics reduced this slightly to 29 percent as seen in the interviewer effects odds ratio. When assessing compliance as measured through reporting on the first personal visit, 26 percent of the variation was attributable to interviewers, which decreased to 9 percent when incorporating the interviewer and case characteristics.

Table 2 shows the results from the multilevel models predicting compliance with this new task. One key finding in these models is that, when presented with more opportunities to comply with this new task, interviewers are more likely to do so. The instrument brings these questions on path every time the interviewer launches the CHI until they are completed. Three covariates in the model show that while interviewers may not record the observations on the first personal visit attempt, they will eventually comply and record the observations during a later contact attempt. Cases started during the first week in the interview period, situations of denied access, and an above average number of contact attempts for the sample unit all resulted in higher odds of recording the observation overall, but lower odds of recording the observation on the first personal visit attempt.

Table 2. Estimated Odds Ratios, Standard Errors, and Variance Components for Multilevel Models Predicting Interviewer Compliance at the NOI Record Level.

	Observations Present	Recorded on 1st Personal Visit[2]
Fixed Effects		
Interviewer Characteristics		
Education (*Ref*=High School)		
Some College	1.17	0.84
Bachelors +	0.99	0.89
Female	0.94	0.97
Years Census Exp[1]	0.99	1.01
Years NHIS Exp[1]	1.00	0.97
Supervisor	0.68	0.93
Intermittent Employee	1.06	1.20
Caseload[1]	1.02	0.88
Multiple Surveys	0.81	1.12
Case Characteristics		
Started Week 1	1.59***	0.87*
Denied Access	5.98***	0.64***
Vacant	0.76**	0.80**
Interim Refusal	3.42***	1.01
Other Eligible Noninterviews	0.49***	0.99
Contact Attempts[1]	1.31***	0.93***
Multi-unit	0.44***	0.98
Urban	2.49***	1.36***
Random Effects		
Interviewer Effects	1.29	1.09
Interviewer ICC	0.33	0.26
Cases	30,629	29,201
Interviewers	858	849
Avg N/group	35.6	34.3

Source: U.S. Census Bureau, National Health Interview Survey, January-April 2013.
*p-value≤0.05, **p-value≤0.01, ***p-value≤0.001
[1] Values centered to the mean for ease of interpretation.
[2] The observations must be present to be included in this model.

Higher odds of complying with both measures occurred when the sample unit was in an urban rather than rural setting, and lower odds of complying when the sample unit was vacant. Interviewers were less likely to record the observations for sample units within multi-unit buildings, though both the question wording and the training indicated they should make the observations regarding the building within which the sample unit resides if they are able to observe the building.[5] When a sample unit refused participation in the survey, interviewers were more likely to record the observation, which may be the result

[5] Some observations, however, are specific to the sample unit and cannot be observed without access to the sample unit itself.

of additional contact attempts. Neither interim refusals nor multi-unit structures significantly predicted the odds of complying with the first personal visit protocol.

During debriefing sessions with interviewers, they voiced concerns regarding observations pertaining to the residents within the sample unit when the unit was obviously vacant, explaining the significantly lower odds of compliance in these situations. We know from speaking with interviewers that situations captured by the other eligible noninterview category more frequently occur in rural settings, which may explain the significance of the urban indicator as well as the noninterview covariate. The latter could be situations where they could not access the sample unit for reasons like snow-covered roads, which would be a legitimate reason to *not* record the observations on the first personal visit attempt.

Overall, interviewer compliance with this new task was high, more so with the overall task itself than the first personal visit protocol. The direction and magnitude of the case characteristic odds ratios can inform future training decisions, and it is beneficial to know that interviewer characteristics do not contribute to variation in compliance with recording the observations.

3.2 Observation Variation

The usefulness of interviewer observations for nonresponse adjustments and adaptive design modeling require independence from contact during the attempt when the observations are recorded (Groves, 2006; Kreuter et al, 2010; Peytchev and Olson, 2007; West, 2013). Table 3 displays the reported presence of the dichotomous observations and the mean for the scaled observations—address condition and income—by contact outcome.[6] This is first displayed overall, then by the outcome of the contact attempt during which the observation was recorded—contact with a sample unit member or noncontact.

Almost half of the observations did not vary based on making contact or not—graffiti, bars on windows, multiple door locks, indicators of disabled residents, presence of an adult bicycle, and household income. The wording of these particular observations focuses more on the presence or absence of specific indicators (facts), and not asking interviewers to make judgments or estimations (inferences), with the exception of the income measure. Interviewers were asked to assign the sample unit's income to the top, middle, or bottom third of the local population based on observation and knowledge of the area. All of the other observations that did not vary based on contact were worded in a manner that asked interviewers to state whether the phenomena of interest was present.

The observations that differed by outcome are those more open to interpretation, and more specifically ask for inferences of the residents within the sample unit. Interviewers were more likely to report the presence of the following when making contact: children, smoking, speaking a language other than English, and residents aged 65 and over. All of these measures would be more apparent if speaking directly to a sample unit member and then become the reporting of facts instead of inferences. When making contact, interviewers rated the condition of the sample unit higher than when recording the

[6] Address condition is reported on a Likert scale, with 1 representing "poor" and 5 representing "excellent." Interviewers identified the sample unit as having income in the bottom third, middle third, or top third of the local population.

observations on an attempt when not making contact. Interviewers were less likely to report the following when making contact: access barriers, well-tended yards, damaged walls, and employed residents. Overall, we saw inferentially based interviewer observations differed significantly when making contact with a sample unit member during the attempt when recording the observations, echoing the interviewers' comments during debriefing sessions indicating comfort with reporting facts but not making inferences.

Table 3. Interviewer Observation Proportions and Means

Interviewer Observation	Overall	Noncontact	Contact	
Graffiti	0.04	0.04	0.04	
Condition of Sample Unit†	3.78	3.73	3.87	*
Buzzer/keycode/doorman	0.16	0.18	0.14	*
Well-tended yard/garden	0.56	0.57	0.32	*
Peeling paint/damaged walls	0.14	0.16	0.12	*
Bars on Windows	0.06	0.06	0.06	
Multiple door locks	0.04	0.04	0.04	
Presence of children <6	0.11	0.08	0.17	*
Wheelchair ramp/disabled	0.03	0.03	0.04	
Adult Bicycle	0.03	0.02	0.04	
Evidence of Smokers	0.07	0.05	0.10	*
Household Income bracket†	1.80	1.80	1.80	
Employed adult(s)	0.72	0.78	0.72	*
Language other than English	0.14	0.11	0.20	*
Household aged 65+	0.12	0.07	0.20	*
N	29,201	11,412	17,789	

Source: U.S. Census Bureau, National Health Interview Survey, January-April 2013.
* indicates statistically significant difference at the $p \leq 0.05$ level when making contact for that observation.
† indicates reported value is a mean instead of proportion.

3.3 Observations as Predictive Measures

The first column in Table 4 presents the negative binomial regression coefficients (labeled as 'NBR') for a model regressing the number of contact attempts on interviewer and case characteristics as well as the individual observations and applicable interaction terms. When interviewers reported the presence of disability indicators, cases were likely to require significantly fewer contact attempts. The better the reported condition of the sample unit and the higher the perceived income of the household, the fewer the number of contact attempts. When interviewers perceived residents of the sample unit as employed, the interviewers were likely to make significantly more contact attempts.

Though not shown here, we ran models stratified by contact (with a sample unit member versus noncontact) to determine the interaction between making contact and the utility of the observations for predicting level of effort. If either the direction or magnitude of the relationship between the observation and the level of effort measure changed based on contact, we included an interaction term in our final model. Overall, only two of the observations were not dependent upon making contact—at least one of the adults in the sample unit is employed and evidence of a disabled sample unit member. Evidence of smoking and households aged 65 and over emerged as statistically significant only when recorded during a visit where the interviewer made contact with a sample unit member.

The second model in Table 4 displays the odds ratios for regressing whether the case is an interim refusal or not on interviewer and case characteristics as well as the individual observations and applicable interaction terms. Reporting well-tended yards, evidence of disabled household members, and households aged 65 and over positively and significantly predicted interim refusals. Turning to the interactions, when interviewers recorded the observations but did not make contact with a sample unit member, both income and employment positively and significantly predicted interim refusals. The presence of an adult bicycle, employed sample unit members, and households with at least one member who speaks a language other than English significantly predicted the likelihood of interim refusals when interviewers recorded the observations during an attempt that resulted in contact with a sample unit member. For each of these three observations, when interviewers made contact, sample units were approximately 3.5 times more likely to be interim refusals.[7] The condition of the sample unit, barriers to access, damaged walls, barred windows, multiple door locks, presence of children, and indicators of smoking did not significantly predict the odds of a sample unit being an interim refusal.

Table 4. Model Parameters Predicting Number of Contact Attempts and Presence of Interim Refusals

	# Contact Attempts	Interim Refusal
Interviewer Characteristics	NBR Coefficients	Odds Ratios
Education (*Ref*=High School)		
Some College	-0.03	0.97
Bachelors +	0.05***	1.04
Female	0.01	0.86***
Years Census Exp	-0.01*	1.01
Years NHIS Exp	0.00	0.98*
Supervisor	0.03	1.22**
Intermittent Employee	-0.10***	1.15***
Caseload (above mean)	-0.10***	0.83***
Multiple Surveys	-0.08***	0.95
Case Characteristics		
Started Week 1	0.09***	1.22***
Denied Access	0.14***	1.05
Vacant	-0.50***	0.13***
Interim Refusal	0.53***	N/A
Contact Attempts	N/A	1.18***
Other Eligible Noninterviews	0.43***	1.38***
Multi-unit	0.05**	0.77***
Urban	0.14***	1.04
Interviewer Observations		
Graffiti	-0.02	0.92

[7] Though not shown, the exponentiation of the odds ratios for contact, the interaction, and the interviewer observation were applied to calculate the total effect, which was then converted back to an odds ratio.

Condition of Sample Unit	-0.04***	1.00
Buzzer/keycode/doorman	-0.07	1.06
Well-tended yard/garden	-0.02	1.10*
Peeling paint/damaged walls	-0.03	0.98
Bars on Windows	0.00	1.11
Multiple door locks	0.02	1.12
Presence of children <6	0.01	0.94
Wheelchair ramp/disabled	-0.10**	0.79*
Adult Bicycle	-0.02	1.11
Evidence of Smokers	-0.01	0.91
Household Income bracket	-0.03*	1.12**
Employed adult(s)	0.20***	1.31***
Language other than English	0.00	1.07
Household aged 65+	0.00	1.22***
Made Contact	-0.44***	3.34***
Interactions with Contact		
Damaged Walls	-0.01	N/A
Children <6	0.06	N/A
Adult Bicycle	N/A	0.63*
Smoker	-0.14**	0.86
Income	N/A	1.00
Employment	N/A	0.67***
Language	0.06	0.78**
Older Household	-0.19***	N/A

Source: U.S. Census Bureau, National Health Interview Survey, January-April 2013.
N=29,201 NOI Records; *p-value≤0.05, **p-value≤0.01, ***p-value≤0.001

Overall, four of the interviewer observations were not statistically significant predictors of either measure of effort—graffiti, barred windows, multiple door locks, and indicators of children. Four of the observations were only predictive of either measure of level of effort when recorded after making contact with a sample unit member—damaged walls, indicators of smoking, presence of an adult bicycle, and speaking a language other than English. Seven of the observations predicted level of effort regardless of making contact when recording the observations. Income, address condition, and access barriers were all predictive of the number of contact attempts for a case. Age and well-tended yards were both predictive of an interim refusal. Evidence a sample unit member may be employed significantly predicted both measures.

4. Discussion

The purpose of this research was to assess interviewer observations collected during the fielding of a household survey, looking at interviewer compliance, as well as the variation in and the utility of these measures. Interviewer compliance with this new task was high—they are recording the observations, but not necessarily on the first personal visit. A considerable amount of variation exists between interviewers when making these observations, though it is only the case characteristics and *not* the interviewer characteristics that significantly contribute to reducing this variation across interviewers.

Looking at the variation of the individual interviewer observations, our findings support feedback from interviewers, specifically with respect to making contact with a sample unit member when recording the observations. We found evidence to suggest differences in reporting inferentially based observations when making contact with a sample unit member. In addition, reporting the absence of person-based characteristics was more prevalent than reporting the presence of these characteristics when interviewers did not make contact. This variation may be the result of making contact, or simply the difference between responding and nonresponding units (refer to Erdman and Dahlhamer, 2013, for the relationship between the observations and a sample unit's propensity to respond).

If interviewers make contact with a sample unit member, they are less likely to report barriers to access and employed residents. This reflects comments heard during debriefing sessions with interviewers. Though instructed to make the observations of the building within which the sample unit resides, we heard interviewer reluctance to record observations when access barriers were present.

One clear assumption of interviewers is that sample unit members are not home during the day because they are working. We saw this demonstrated in the variation of reporting sample unit members as employed based on making contact when recording the observations. This is also consistent with the findings of Sinibaldi and colleagues (2013), where interviewer accuracy for reporting employment was 93 percent, though the accuracy was lower when the sample unit members refused to participate in the survey. Interviewer assessment of sample unit member employment status predicted both measures of level of effort—number of contact attempts and being an interim refusal—regardless of making contact when recording the observations. This suggests further research is necessary to assess accuracy of this estimation once the survey response data are available.

Interviewers were more likely to report the presence of children, evidence of smoking, speaking a language other than English, and households aged 65 and over when making contact. While people who smoke may not leave evidence outside the residence, just as those with children may not leave toys in the yard, if an interviewer makes contact with a resident who smokes or has children, the evidence is far more prevalent. This is similarly the case with regard to the age of the sample unit members and the language(s) spoken in the home. These findings align with previous research that found interviewers have more difficulty discerning the presence of children as opposed to the absence of children (Landis and Koch, 1977; West, 2013).

We found that the interviewer observations predict both measures of level of effort—the number of contact attempts and interim refusals. For example, employment and outward evidence of disabled residents were both statistically significant predictors of both level of effort measures regardless of making contact with a sample unit member when recording the observations. We found income, address condition, and access barriers predict the number of contact attempts. The age of the sample unit members and having well-tended yards predict the odds of a sample unit being an interim refusal. Providing this information to managers on a daily basis could reduce the level of effort necessary to complete a case, specifically with respect to case reassignment. Cases with an increased likelihood of being an interim refusal could be reassigned to a refusal conversion expert earlier in the interview period, thereby reducing the number of contact attempts.

While this evaluation of the interviewer observations is generally positive, additional work is necessary. Once the survey response data are available, the accuracy of these observations can be more directly assessed. Additionally, assessing the interviewer observations against the survey response data determines whether the observations are predictive of key survey estimates, which is the second component of using observations for both adaptive design models and nonresponse adjustments (Blom et al 2011; Durrant et al 2012; Groves and Heeringa, 2006; Kreuter et al 2010; Cases-Cordero et al, 2013; West, 2013).

This research demonstrated the high level of compliance with the new task of recording observations, while drawing attention to the variation across interviewers and observations. Making contact with a sample unit member when recording the observations specifically alters the recording of observations that may have arbitrary observable indicators from outside the sample unit (the presence of children, smoking sample unit members, language spoken, and age of the residents), requiring interviews to make inferences with little observable information. Several of the observations can predict at least one measure of the level of effort, though some of these also vary based on making contact when recording the observations.

The results from this and other analyses at the Census Bureau are being discussed to make modifications to question wording and interviewer training. Research is ongoing with respect to the utility of these observations. Overall, the high compliance rates and our initial findings are promising for use of interviewer observations in adaptive design.

References

Agresti, A. and Finlay, B. (2009). Statistical Methods for the Social Sciences. Pearson Prentice Hall, Michigan.

Atrostic, B., Bates, N., Burt, G., and Silberstein, A. (2001). Nonresponse in U.S. Government Household Surveys: Consistent measures, recent trends, and new insights. *Journal of Official Statistics*, 17(2):209-226.

Axinn, William, Cynthia F. Link, and Robert M. Groves. 2011. "Responsive Survey Design, Demographic Data Collection, and Models of Demographic Behavior." *Demography*, 48(3): 1127-1149.

Bates, N., J. Dahlhamer, and E. Singer (2008). “Privacy concerns, too busy, or just not interested: Using doorstep concerns to predict survey nonresponse.” *Journal of Official Statistics* 24 (4): 591-612.

Bethlehem, J. (2002). Weighting nonresponse adjustments based on auxiliary information. In R. Groves, D. Dillman, J. Eltinge, and R. Little (eds.) Survey Nonresponse. Wiley, New York.

Biemer, P., Chen, P., and Wang, K. (2012). Using Level of Effort Paradata in Nonresponse Adjustments with Application to Field Surveys. *Journal of the Royal Statistical Society*, 176, 147-168.

Blom, A, de Leeuw, E. and Hox, J(2011) Interviewer effects on nonresponse in the European Social Survey. *Journal of Official Statistics*, 27, 359-377.

Casas-Cordero, C., Kreuter, F., Wang, Y, and Babey, S. (2013). Assessing the measurement error porperties of interviewer observations of neighbourhood characteristics. *Journal of the Royal Statistical Society*, 176, 227-249.

Dahlhamer, J. M. (2012). New Observation Questions. Presentation given at the 2013 NHIS Centralized Refresher Training and Conference, Dallas, TX, December 3-6.

Durrant, G., D'Arrigo, J. and Steele, F. (2011) Using Paradata to predict best time of contact, conditioning on household an interviewer influences. *Journal of Official Statistics*, 174, 1029-1049.

Erdman, C. and Dahlhamer, J. (2013) Evaluating Interviewer Observations in the National Health Interview Survey: Associations with response propensity. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*. Montreal, CN.

Groves, R. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5), 646-675.

Groves, R. and Heeringa, S. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3): 439-457.

Kreuter, F., K. Olson, J. Wagner, T. Yan, T.M. Ezzati-Rice, C. Casas-Cordero, M. Lemey, A. Peytchev, R.M. Groves, T.E. Raghunathan. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Nonresponse: Examples from multiple surveys. *Journal of the Royal Statistical Association*, 173(2), 389-407.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Lessler, J. and Kalsbeck, W. (1992). Nonresponse: dealing with the problem. In Nonsampling Errors in Surveys, New York: Wiley-Interscience.

Little, R. and Vartivarian, S. (2005). Does Weighting for Nonresponse increase the variance of survey means? *Survey Methodology*, 31:161-168.

Lynn, P. (2003). PEDAKSI: methodology for collecting data about survey nonrespondents. *Quality and Quantity*, 37, 239-261.

Maitland, A., Casas Cordero, C., and Kreuter, F. (2008). An Exploration into the Use of Paradata for Nonresponse Adjustment in a Health Survey. Pp. 2250-2255 in *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods.*

Miller, P., Bates, N., Dahlhamer, J., and Gindi, R. (2013). Developing Interviewer Observations of the Neighborhood and Sample Unit for the National Health Interview Survey. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*. Montreal, CN.

Mohl, C. and LaFlamme, F. (2007). Research and responsive design options for survey data collection by Statistics Canada. American Statistical Associations's Annual Joint Statistical Meeting, August, 2007. Session of Survey Research Methods. Salt Lake City, UT.

Peytchev, A. and Olson, K. (2007). Using Interviewer Observations to Improve Nonresponse Adjustments: NES 2004. American Statistical Association Joint Statistical Annual Meeting. *Session on Survey Research Methods*. Salt Lake City, UT.

Rabe-Hesketh, S., & Skrondal, A. (2005). *Multilevel and longitudinal modeling using STATA*. College Station, TX: STATA Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Sinibaldi, J., Durrant, G., and Kreuter, F. (2013). Evaluating the measurement error of interviewer observed paradata. *Public Opinion Quarterly*, 77, 173-193.

West, B. (2013). An examination of the quality and utility of interviewer observation in the National Survey of Family Growth. *Journal of the Royal Statistical Society*, 176(1), 211-225.

West, B. and Kreuter, F. (2013). Factors affecting the accuracy of interviewer observations: Evidence from the National Survey of Family Growth. *Public Opinion Quarterly*, 77(2), 522-548.